AN INTRODUCTION TO THE
THEORY OF STATISTICS.

AN INTRODUCTION TO THE
THEORY OF STATISTICS

BY

G. UDNY YULE, C.B.E., M.A., F.R.S.,

FELLOW OF ST JOHN'S COLLEGE, AND FORMERLY READER IN
STATISTICS, CAMBRIDGE; HONORARY VICE-PRESIDENT
OF THE ROYAL STATISTICAL SOCIETY

AND

M. G. KENDALL, M.A.,

FORMERLY MATHEMATICAL SCHOLAR OF ST JOHN'S COLLEGE, CAMBRIDGE;
FELLOW OF THE ROYAL STATISTICAL SOCIETY.

With 55 Diagrams and 4 Folding Plates.

ELEVENTH EDITION, REVISED THROUGHOUT AND RE-SET.

ABRIDGED PREFACE TO THE FIRST EDITION.

THE following chapters are based on the courses of instruction given during my tenure of the Newmarch Lectureship in Statistics at University College, London, in the sessions 1902–1909. The variety of illustrations and examples has, however, been increased to render the book more suitable for the use of biologists and others besides those interested in economic and vital statistics, and some of the more difficult parts of the subject have been treated in greater detail than was possible in a sessional course of some thirty lectures. To enable the student to proceed further with the subject, fairly detailed lists of references to the original memoirs have been given: exercises have also been added for the benefit, more especially, of the student who is working without the assistance of a teacher.

The volume represents an attempt to work out a systematic introductory course on statistical methods—the methods available for discussing, as distinct from collecting, statistical data—suited to those who possess only a limited knowledge of mathematics: an acquaintance with algebra up to the binomial theorem, together with such elements of co-ordinate geometry as are now generally included therewith, is all that is assumed. I hope that it may prove of some service to the students of the diverse sciences in which statistical methods are now employed.

G. U. Y.

December 1910.

.

PREFACE TO THE ELEVENTH EDITION.

The "*Introduction to the Theory of Statistics*" having completed five-and-twenty years of life, it was decided that the time had come when a complete revision should be made. This, I felt, I could not personally undertake: it was clearly a task for a younger man, more in touch with recent literature and less affected by the prejudices of age in favour of the old and the familiar.

Mr Kendall undertook the task not merely with willingness but with enthusiasm. I read his typescript, but to him is primarily and almost solely due the credit for suggesting the general lines of the revision, and for carrying out the agreed suggestions: the only new chapter for which I am directly responsible is Chapter 24 on Interpolation and Graduation, based on a few lectures sometimes included in former courses.

I hope that in its new form the book may long continue to be of service to further generations of students.

G. UDNY YULE.

CAMBRIDGE,
July 1937.

In the revision undertaken for this edition, apart from some substitution of new numerical illustrations for old, very little of the material appearing in earlier editions has been deleted. A few minor alterations have been made—the matter formerly included in supplements has been incorporated in the text, and there has been some rearrangement—but the major changes are almost entirely in the form of additions. Of these, the most important are several new chapters on Sampling, including an introductory chapter on Small Samples. Chapters have also been added on Moments and Measures of Skewness and Kurtosis, and on Simple Curve Fitting by the Method of Least Squares. Mr Yule has contributed a new chapter on Interpolation and Graduation. For the first time Tables of the various functions commonly required in statistical work have been assembled at the end of the book. Throughout the preparation of this new material I have had the benefit of Mr Yule's encouragement, criticism and advice.

The complete revision has presented the opportunity of issuing the book in new form, and it is hoped that the larger page and type will be

found an improvement.   A more distinctive system of paragraph number-
ing and paragraph headings has been introduced.   Some further Exercises
have also been added.

 •   Notwithstanding the mathematical character of recent developments
in statistical theory, an attempt has been made to keep within the limits
laid down by Mr Yule for earlier editions of this book in regard to the
knowledge of mathematics required by its readers.   In one or two places
it has been necessary to introduce the notation of the integral calculus,
but this has been accompanied by explanations in terms of geometrical
ideas.

It is a pleasure to record Mr Yule's and my indebtedness to " Student "
and the proprietors of *Metron* for permission to reproduce a slightly
condensed version of the former's tables of the *t*-integral; and to R. A.
Fisher and Messrs Oliver & Boyd for permission to reproduce the tables
of the significance points of the *z*-integral from Professor Fisher's
"*Statistical Methods for Research Workers*."   The tables for the 0·1 per cent.
level of *z* are due to W. E. Deming, Lola S. Deming and C. G. Colcord,
who have also very generously given their consent to the reproduction.

I shall feel indebted to any reader who directs my attention to possible
errors, omissions, ambiguities or obscurities.

<div align="right">M. G. K.</div>

CONTENTS.

LIST OF FOLDING PLATES.

NOTES ON NOTATION AND ON TABLES FOR FACILITATING STATISTICAL WORK.

A. Notation.

The reader is assumed to be familiar with the commoner mathematical signs, *e.g.* those for addition and multiplication. We shall also employ the following symbols, all of which are in general use :—

The Factorial Sign.

The symbol $n!$, read " factorial n," means the number

$$1 \times 2 \times 3 \times \ . \ . \ . \ \times (n-2) \times (n-1) \times n$$

Factorial n is by some writers expressed by the symbol $\lfloor n$, but this notation appears to be falling out of use in favour of $n!$, probably owing to the greater ease with which the latter form can be printed and type-written.

The Combinatorial Sign.

The symbol ^nC_r means the number of ways in which r things can be chosen from n things, *e.g.* $^{52}C_{13}$ is the number of ways in which a hand of cards can be dealt from an ordinary pack of 52 cards.

In most text-books on Algebra it is shown that

$$^nC_r = \frac{n!}{r!(n-r)!} = {}^nC_{(n-r)}$$

The Summation Sign.

The sum of n numbers $x_1, x_2, \ . \ . \ . \ x_n$ is written $\overset{r=n}{\underset{r=1}{S}} (x_r)$, read " sum x_r from one to n," *i.e.*

$$\overset{r=n}{\underset{r=1}{S}} (x_r) = x_1 + x_2 + x_3 + \ . \ . \ . \ + x_{(n-1)} + x_n$$

Where no ambiguity is likely to arise, the suffix r and the limits written above and below S are omitted, *e.g.* the above sum would be written simply $S(x)$, it being understood from the context that the summation extends over the n values.

Many writers use the Greek letter Σ instead of S.

The Greek Alphabet.

As the letters of the Greek alphabet will often be used as symbols, we give for convenience the names of those letters.

Small Letter.	Capital Letter.	Name.	Small Letter.	Capital Letter.	Name.
α	A	alpha	ν	N	nu
β	B	beta	ξ	Ξ	xi
γ	Γ	gamma	o	O	omicron
δ	Δ	delta	π	Π	pi
ϵ	E	epsilon	ρ	P	rho
ζ	Z	zeta	σ, s	Σ	sigma
η	H	eta	τ	T	tau
θ	Θ	theta	υ	Υ	upsilon
ι	I	iota	ϕ	Φ	phi
κ	K	kappa	χ	X	chi (*pron.* ki)
λ	Λ	lambda	ψ	Ψ	psi
μ	M	mu	ω	Ω	omega

B. Calculating Tables.

For heavy arithmetical work a calculating machine is invaluable; but owing to their cost machines are, as a rule, beyond the reach of the student.

For a great deal of simple work, especially work not intended for publication, the student will find a slide rule exceedingly useful : particulars and prices will be found in any instrument-maker's catalogue. For greater exactness in multiplying or dividing, logarithms are almost essential.

If it is desired to avoid logarithms, use may be made of extended multiplication tables. There are a great many of these and some references to different forms are given in the list on pages 524–525.

In addition to general arithmetical tables of this kind, the student will derive invaluable aid from Barlow's " *Tables of Squares, Cubes, Square-roots, Cube-roots, and Reciprocals of All Integral Numbers up to 10,000* " (E. & F. N. Spon, London and New York, price 7s. 6d.), which are useful over a wide range of statistical work.

C. Special Tables of Functions Useful in Statistical Work.

The tables and diagram at the end of this book will cover most of the student's ordinary requirements. Other tables appear in the works cited on page 525. The more advanced student will find it useful to have " *Tables for Statisticians and Biometricians* " (Cambridge University Press, Part 1, price 15s., and Part 2, price 30s.)—particularly Part 1. Research workers will wish to have the tables appended to R. A. Fisher's " *Statistical Methods for Research Workers,*" 6th ed. (Oliver & Boyd, price 15s.).

D. References to the Text.

Each section in the book is distinguished by a number in heavy type consisting of the number of the chapter in which the section occurs prefixed to the number of the section in that chapter and separated from it by a period ; *e.g.* **7.13** means the Thirteenth Section of Chapter 7, and **10.1** refers to the First Section of Chapter 10. The Introduction,

which precedes Chapter 1, is for this purpose regarded as Chapter 0, *e.g.* **0.26** refers to the Twenty-sixth Section of the Introduction. References to sections are given simply by the number of the sections, *e.g.* " We saw in **8.3** " means " We saw in the Third Section of Chapter 8."

Similarly, equations, tables, examples, exercises and diagrams are distinguished first of all with the number of the chapter in which they occur and then, separated by a period, with their serial number within the chapter, *e.g.* " Table **6.7** " refers to the Seventh Table in Chapter 6, and " Equation (**17.8**) " refers to the Eighth Equation of Chapter **17.** These figures are in ordinary type.

This simple notation saves a good deal of unnecessary wording. To facilitate quickness of reference we sometimes give pages as well.

A distinction is drawn between *examples*, which are given in the text for purposes of illustration, and *exercises*, which are set at the end of the chapter for the student to work out for himself.

THEORY OF STATISTICS.

INTRODUCTION.

Number and Measurement.

0.1. Western civilisation is pervaded by ideas of number and measurement. Even the events of our everyday life are inextricably bound up with them. We have only to picture a race which cannot count or measure trying to run the Bank of England, or control the milk market, or even understand the sporting columns of the daily press, to realise how deeply rooted numbers are in the complex activities of the modern world.

0.2. Science itself is particularly indebted to numerical expression. As organised knowledge has increased, the necessity for precision has become greater, and in the formulation of precise statements number and measurement have played a leading part. The desire for quantitative expression was first felt in the physical sciences, but it has now spread into nearly all branches of knowledge. The movement is by no means complete, however, and may be seen at work to-day. As a significant instance we may note that courageous attempts are being made to subject the process of thought itself—that last stronghold of the contentious and the mysterious—to quantitative inquiry.

0.3. Many people, in fact, have been led by their enthusiasm for numerical data to regard knowledge of a non-quantitative kind as hardly deserving the name " knowledge " at all. Towards the close of the nineteenth century it was possible for Lord Kelvin to say : " When you can measure what you are speaking about and express it in numbers you know something about it ; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind." This remark has often been quoted with an approval which it does not altogether deserve it does not, for example, do justice to the work of Darwin and Pasteur, to name only two of Kelvin's contemporaries. But there can be no denying that it expresses a point of view which many people will endorse.

Numerical Data.

0.4. The desire for precision, in fact, leads investigators of all kinds, from the atomic physicist to the business man, to express the facts about that part of the universe which interests them in a quantitative way. Numerical data have come into being not only in the laboratory and the study, but in the counting-house, the sales department, the Board Room and the legislative assembly. It is difficult to see how our society could be organised without them. Where the Jews and the Romans were content

1

with occasional censuses for military or fiscal purposes,[1] the progressive modern state finds itself under the necessity of keeping a close and quantitative eye on all that goes on within or without its frontier. A country which does not do so may be fairly regarded as backward. In a typical phrase Anatole France summed up this point of view when he said of the Chinese : " Tant qu'ils ne se seront pas comptés, ils ne compteront pas "— if they don't count they won't count.

Statistics Concerned with Numerical Data.

0.5. There are certain features of numerical data, no matter in what branch of knowledge they originate, which may call for a special type of scientific method to treat them and elucidate them. This is known as **Statistical Method,** or more briefly, as **Statistics.** It does not, however, embrace the study of numerical data of every kind, and before we attempt a formal definition of its nature and scope, it is necessary to give some words of explanation.

Effects and Causes.

0.6. One of the principal aims of Science is to trace, amidst the tangled complex of the external world, the operation of what are called " laws "— to interpret a multiplicity of natural phenomena in terms of a few fundamental principles. A knowledge of the operation of these laws enables us to talk of " cause " and " effect." The metaphysical problems associated with these words need not detain us, but since in the sequel we shall often use them, it is proper to explain that we adopt them as a convenient way of expressing serviceable and familiar ideas. We need not worry if the atomic physicist says that causation must be rejected. We shall be dealing with the everyday world, where " law " and " cause " have significant and important connotations.

0.7. With this convention, we may say that any physical event, and in particular that described by quantitative data, is produced by the operation of one or more causes. The number of causes which produce any particular effect may be, and usually is, extremely large. For instance, the height of a man is causally linked with his race, his ancestry, his habitation, his diet during youth, his age, his occupation, and at any given moment even with his position and the time of day.

0.8. Experiment, the great weapon of scientific inquiry, derives its power from the ability of the experimenter to replace such complex systems of causation by simple systems in which only one causal circumstance is allowed to vary at a time. This is perhaps an ideal, but it is one which is closely approached with the technique of modern laboratory practice.

0.9. Let us, however, turn to social science, as the parent of the methods termed " statistical," for a moment, and consider its characteristics as compared, say, with physics or chemistry. One characteristic stands out so markedly that attention has been repeatedly directed to it

---

[1] David (II Samuel, 24) numbered the people of Israel and called down a plague by doing so. He counted 800,000 valiant men who drew the sword, and though the text is not entirely clear it seems likely that Divine disapproval was directed against the militaristic purpose of the census, not the census itself. We are told later that 70,000 men died of the resulting pestilence, so it looks as if there was no ban on counting *dead* men.

by " statistical " writers as the source of the peculiar difficulties of their science—*the observer of social facts cannot experiment, but must deal with circumstances as they occur, apart from his control.* The simplification open to the experimenter being impossible, the observer has; in general, to deal with highly complicated cases of multiple causation—cases in which a given result may be due to any one of a number of alternative causes or to a number of different causes acting conjointly.

0.10. A little consideration will show that this is also characteristic of observations in other fields. The meteorologist, for example, is in almost precisely the same position as the student of social science. He can experiment on minor points, but the records of the barometer, thermometer and rain gauge have to be treated as they stand. With the biologist, matters are somewhat better. He can and does apply experimental methods to a very large extent, but frequently cannot approximate closely to the experimental ideal ; the internal circumstances of animals and plants too easily evade complete control. Hence a large field (notably the study of variation and heredity) is left in which methods of experiment have to be supplemented by other methods. The physicist and chemist, finally, stand at the other extremity of the scale. Theirs are the sciences in which experiment has been brought to its greatest perfection. But even so, there is still scope for the application of statistical treatment in these sciences. The methods available for eliminating the effect of disturbing circumstances, though continually improved, are not, and cannot be, absolutely perfect. The observer himself, as well as the observing instrument, is a source of error; the effects of changes of temperature, or of moisture, or pressure, and draughts, vibration, etc., cannot be completely eliminated.

0.11. It is with data affected by numerous causes that Statistics is mainly concerned. Experiment seeks to disentangle a complex of causes by removing all but one of them, or rather by concentrating on the study of one and reducing the others as far as circumstances permit to a comparatively small residuum. Statistics, denied this resource, must accept for analysis data subject to the influence of a host of causes, and try to discover from the data themselves which causes are the important ones and how much of the observed effect is due to the operation of each.

Definitions.

0.12. In the light of the foregoing discussion we may accordingly give the following definitions :—

By **Statistics** we mean quantitative data affected to a marked extent by a multiplicity of causes.

By **Statistical Methods** we mean methods specially adapted to the elucidation of quantitative data affected by a multiplicity of causes.

By **Theory of Statistics** or, more briefly, **Statistics** we mean the exposition of statistical methods.

(It will be observed that the same word may be used both for the science and for the raw material on which it works. This dual use gives rise to no confusion in practice, but the distinction is worth bearing in mind.)

Use of " Statistic."

0.13. This is perhaps the appropriate place to remark that there has recently come into use the singular form " statistic." This is the name

given to a particular kind of estimate compiled from observations, usually according to some algebraical formula. In this book we shall rarely, if ever, have occasion to use the term, and we mention it mainly to forewarn the student who may meet the term elsewhere or in further reading. We may also point out that Statistics is not confined to the study of such entities any more than Physics is the study of individual articles of physic.

History of the word " Statistics."

0.14. In their present meaning the words " statistics," " statistician " and " statistical " are less than a century old. They have, however, been in use longer than that, and it is instructive to consider the process by which they have reached their present meaning.

0.15. The words " statist," " statistics," " statistical" appear to be all derived, more or less indirectly, from the Latin *status*, in the sense, acquired in mediæval Latin, of a political *State*.

0.16. The first term is, however, of much earlier date than the two others. The word " statist " is found, for instance, in *Hamlet* (1602)[1], *Cymbeline* (1610 or 1611),[2] and in *Paradise Regained* (1671).[3] The earliest occurrence of the word " statistics " yet noted is in *"The Elements of Universal Erudition,"* by Baron J. F. von Bielfeld, translated by W. Hooper, M.D. (3 vols., London, 1770). One of its chapters is entitled *Statistics*, and contains a definition of the subject as " The science that teaches us what is the political arrangement of all the modern states of the known world." [4] " Statistics " occurs again with a rather wider definition in the preface to " *A Political Survey of the Present State of Europe,*" by E. A. W. Zimmermann,[5] issued in 1787. " It is about forty years ago," says Zimmermann, " that that branch of political knowledge, which has for its object the actual and relative power of the several modern states, the power arising from their natural advantages, the industry and civilisation of their inhabitants, and the wisdom of their governments, has been formed, chiefly by German writers, into a separate science. . . . By the more convenient form it has now received . . . this science, distinguished by the new-coined name of *statistics*, is become a favourite study in Germany " (p. ii); and the adjective is also given (p. v): " To the several articles contained in this work, some respectable *statistical* writers have added a view of the principal epochas of the history of each country."

0.17. Within the next few years the words were adopted by several writers, notably by Sir John Sinclair, the editor and organiser of the first *"Statistical Account of Scotland,"* [6] to whom, indeed, their introduction has been frequently ascribed. In the circular letter to the Clergy of the Church of Scotland issued in May 1790,[7] he states that in Germany " ' Statistical Inquiries,' as they are called, have been carried to a very great extent," and adds an explanatory footnote to the phrase " Statistical Inquiries "—

---

[1] Act 5, sc. 2.   [2] Act 2, sc. 4.   [3] Bk. 4.

[4] We cite from Dr W. F. Willcox, *Quarterly Publications of the American Statistical Association*, vol. 14, 1914, p. 287.

[5] Zimmermann's work appears to have been written in English, though he was a German, Professor of Natural Philosophy at Brunswick.

[6] Twenty-one vols., 1791–99.

[7] *"Statistical Account,"* vol. 20, Appendix to "The History of the Origin and Progress . . ." given at the end of the volume.

" or inquiries respecting the population, the political circumstances, the productions of a country, and other matters of state." In the " History of the Origin and Progress " [1] of the work, he tells us, " Many people were at first surprised at my using the new words, *Statistics* and *Statistical*, as it was supposed that some term in our own language might have expressed the same meaning. But in the course of a very extensive tour, through the northern parts of Europe, which I happened to take in 1786, I found that in Germany they were engaged in a species of political inquiry, to which they had given the name of *Statistics* ; [2] . . . as I thought that a new word might attract more public attention, I resolved on adopting it, and I hope that it is now completely naturalised and incorporated with our language." This hope was certainly justified, but the meaning of the word underwent rapid development during the half-century or so following its introduction.

0.18. " Statistics " (statistik), as the term was used by German writers of the eighteenth century, by Zimmermann and by Sir John Sinclair, meant simply the exposition of the noteworthy characteristics of a state, the mode of exposition being—almost inevitably at that time —preponderantly verbal. The conciseness and definite character of numerical data were recognised at a comparatively early period—more particularly by English writers—but trustworthy figures were scarce. After the commencement of the nineteenth century, however, the growth of official data was continuous, and numerical statements, accordingly, began more and more to displace the verbal descriptions of earlier days. " Statistics " thus insensibly acquired a narrower signification, viz. the exposition of the characteristics of a State by *numerical* methods. It is difficult to say at what epoch the word came definitely to bear this quantitative meaning, but the transition appears to have been only half accomplished even after the foundation of the Royal Statistical Society in 1834. The articles in the first volume of the *Journal*, issued in 1838–39, are for the most part of a numerical character, but the official definition has no reference to method. " Statistics," we read, " may be said, in the words of the prospectus of this Society, to be the ascertaining and bringing together of those facts which are calculated to illustrate the condition and prospects of society." It is, however, admitted that " the statist commonly prefers to employ figures and tabular exhibitions."

0.19. Once the first change of meaning was accomplished, further changes followed. From the name of a science, the word was transferred to those series of figures on which it operated, so that one spoke of vital statistics, poor-law statistics, and so on. It was then applied to the similar numerical data which occurred in other sciences, such as anthropology and meteorology. By the end of the nineteenth century we find " statistics of mental characteristics in man," " statistics of children under the headings bright- average–dull," and even " an examination of the characteristics of the Virgilian hexameter with statistics." The development of the meaning of the adjective " statistical " and the noun " statistician " was naturally similar.

---

[1] *Loc. cit.*, p. xiii.
[2] The "*Abriss der Staatswissenschaft der Europäischen Reiche*" (1749) of Gottfried Achenwall, Professor of Politics at Göttingen, is the volume in which the word "statistik" appears to be first employed, but the adjective "statisticus" occurs at a somewhat earlier date in works written in Latin.

0.20. Perhaps the most abstract use of the word occurs in the theory of thermodynamics, wherein one speaks of *entropy* as *proportional to the logarithm of the statistical probability of the universe*—a definition which no statesman would be unwilling to admit to lie completely outside his purview. But it is unnecessary to multiply instances to show that the word " statistics " is now entirely divorced from " matters of State."

The Theory of Statistics.

0.21. The *theory* of statistics as a distinct branch of scientific method is of comparatively recent growth. Its roots may be traced in the work of Laplace and Gauss on the theory of errors of observation, but the study itself did not begin to flourish until the last quarter of the nineteenth century. Under the influence of Galton and Karl Pearson remarkable progress was made, and the foundations of the subject were laid in the next thirty years—as it has turned out, very securely. The subject has not, however, yet reached a stage whereat a cut-and-dried exposition of its methods can be given. Research, particularly into the mathematical theory of statistics, is rapidly proceeding, and fresh discoveries are being made with a rapidity which makes it difficult to keep pace with them. It may, however, help the student to appreciate the work of later chapters if we sketch in brief general terms the field of statistical theory as it now exists.

The Collection of Data.

0.22. The first question which the statistician has to consider is the collection and assembling of his data. In many fields, such as economics and sociology, he cannot prepare the data himself but has to get what he can from such sources as official statistics, which are usually prepared with an object differing from his own. Such information is therefore rarely all that one could wish. Investigator A, studying the sugar market, finds that the official figures run cane and beet sugar together. Investigator B, wanting to compare prices over a period of years, finds that during the war period 1914–18 there is a gap in the information. Investigator C, wishing to study poverty, has to content himself with indirect figures such as those of poor-law relief and unemployment. But however incomplete the data may be, and however tangentially pertinent to his inquiry, the investigator must take what he can get and be thankful.

0.23. In other cases, and particularly in meteorology, biology and psychology, he can produce his own data or borrow those of other investigators similarly engaged. He does not merely take his figures from some source or other ; he is instrumental in their production, and within limits can control their nature so as to bring them to bear directly on his inquiry.

It might be thought that the only qualities required for such work are an ability to count or measure and a reasonable care. But this is not so. Once outside the laboratory the investigator is beset with a swarm of practical difficulties. We might illustrate the point by referring to the troubles of an investigator who wished to find out how many dairy cows there were in a certain parish. He took the simplest course and went to all the farms in the parish and asked the occupier how many cows he had. Farmer A said that he had fifteen, but had sold eight and was waiting for the buyer to come and fetch them. Farmer B had " about twenty."

Farmer C obviously could not be bothered and said the first figure which came into his head ; and so on.   It is clear that the result of such an inquiry would be to give a quite illusory figure.

0.24.   A full discussion of such matters lies outside the scope of this book, but we have given them more than a passing mention in order to introduce one very necessary caution.

The reliability of data must always be examined before any attempt is made to base conclusions on them.   This is true of all data, but particularly so of numerical data, which do not carry their quality written large upon them.   It is a waste of time to apply the refined theoretical methods of statistics to data which are suspect from the beginning.

The Treatment of Data.

0.25.   Having obtained his data and satisfied himself that they are reliable enough to permit him to proceed, the statistician must then " lick them into shape."   He must decide on some form of arrangement and presentation, reduce them to a convenient scale of units, and so on ; in short, he must work on his raw material until it is ready for the application of his prepared tools.

0.26.   The only process of treatment to which attention need be called is that of condensation.   The mind is incapable of grasping the significance of a large mass of figures.   If, therefore, the quantity of data available is of any size, some process of condensation is necessary to enable the mind to appreciate the picture which the data represent.

Suppose, for instance, we are discussing the stature of a thousand men, and have as data the height of each man to the nearest inch.   Our raw material then consists of a thousand sets of figures ranging from four feet to seven feet, or thereabouts.   Only the supermind could look over these figures and grasp their essentials.   Nor would the position be met by rearranging the figures in order of magnitude.   To get a clear picture of the situation some condensation is necessary, and in this case it can be carried out easily by grouping together all the men whose heights lie in a certain range, say of three inches.   Our total range of three feet is then replaced by twelve sub-ranges, each of three inches, and we may summarise the data by giving the numbers of men who fall into the twelve sub-ranges. In short, we have replaced our original thousand figures by twelve.

0.27.   It will be clear that in so doing we have sacrificed a certain amount of information.   Twelve figures cannot possibly tell us as much as a thousand.   It may very well be, however, that the information in the twelve is all that we require ; the lost information may be irrelevant to the inquiry.   Such a case would happen if we wanted to know, to an inch or so, what was the height exhibited by the greatest number of men.

0.28.   The process of condensation thus sacrifices information but gives us instead a very necessary clarity and adaptability for manipulation. How far the process is carried in any particular case will depend on how far the disadvantages of the sacrifice are offset by the advantages of the clarity.

Summarising and Descriptive Statistics.

0.29.   The process of summarising which we have just described may be carried a great deal further, and leads to a branch of theory which has very important practical applications.

The reader is probably familiar already with the idea of an " *average value*," and with its use in compressing into a single number the results of a series of observations. Such quantities are, in fact, the result of summarising to the greatest possible extent ; they are summaries in which the statistician has distilled the information of a diffuse mass of figures into a single drop, so to speak.

0.30. There is a wide demand for such summarising numbers, and a good deal of this book will be devoted to considering them from one aspect or another. They give a convenient bird's-eye view of what is sometimes a complex and confusing whole. Special sciences have evolved special quantities of this type to meet their own needs. For instance, the economist has invented various kinds of index numbers to express in a shorthand way complicated changes in prices ; and the psychologist has devised coefficients to express the reactions of an individual mind to a sequence of tests.

0.31. The remarks we made in **0.27** and **0.28** apply here with additional force. It must never be forgotten that in summarising we omit. Part of the statistician's task is to see that we do not omit too much.

0.32. The problem of describing a complicated set of data in as few terms as possible is facilitated by the use of mathematical functions. Suppose, for instance, that in the thousand men of **0.26** we assumed that the number of men (y) of height x inches varied as the square of x— frankly a most improbable result, but one which will serve for the purposes of illustration. Then we may describe the data completely by an equation of the form

$$y = ax^2$$

where a is a constant to be determined from the data. Knowing a, we can find the number of men of any given height.

0.33. In this case it rather looks as if we have condensed all the information into a single number a without losing any of it. But that is not so. What we have done is to replace the set of a thousand figures by an assumption about their nature. We have lost none of the information because we assumed, in using the equation, that the information was of a type known to us already.

0.34. It is found in practice that many sets of data may be very conveniently expressed by mathematical functions. The question as to which functions are the most suitable for purposes of description leads to some interesting theory, some of which will be dealt with later and some of which is of an advanced character lying outside the scope of an Introduction to the Theory of Statistics. Such functions are particularly helpful in the theory of sampling.

Analysis of Data.

0.35. When the statistician has arranged and compressed his data into a suitable form, or decided on the functions and evaluated the quantities which he has chosen to describe them, the first stage of his inquiry is finished. It may be that he would wish to take it no further ; for instance, if he is preparing an index number for the economist he may wish to hand over the number to that person without comment, for him to make such use of it as he thinks fit. More frequently, however, he has prepared the

data for his own use as a statistician. He then proceeds to the next stage, that of analysis and elucidation of the causal system which gave rise to them.

0.36. The methods for such purposes are very numerous. In this brief review we need only point out the importance of the investigation of *relationship*, the theory of which bulks very large in statistical literature. If two events are related there is usually, though not always, some causal nexus between them. The problems of the investigation of relationship between phenomena lead to the theory of dependence, contingency and correlation, and the formulation of various coefficients to measure the extent to which one set of events depends upon another.

Sampling.

0.37. When we wish to discuss the properties of an aggregate we may be prevented by practical or theoretical reasons from examining every single member of it. For example, in considering the stature of the male inhabitants of the United Kingdom we cannot measure every man, because of the time and trouble involved ; and in considering the scores of a roulette wheel we cannot examine every score, because the number is practically infinite and observations can be continued as long as the wheel lasts.

0.38. We do not despair, nevertheless, of being able to gain some knowledge of the aggregate. Where we cannot take the whole we do the best we can and try to obtain a selection of members. This selection is called a *sample*.

0.39. It is clear that a sample will not tell us everything about the parent aggregate from which it is derived. Nevertheless, most people have a feeling, and we shall see later in this book that under certain conditions the feeling is a justifiable one, that the sample will give us some information about the parent. Values calculated from the sample may be taken to be estimates of values in the parent, to a degree of approximation which becomes closer as the sample gets larger ; and even where the sample is small we can sometimes draw inferences of a general nature about the parent.

0.40. We are rarely, if ever, able to reason from the sample to the parent with the categorical certainty of a mathematical proof. Our inferences will usually be expressed in terms of probabilities. Moreover, we shall find it much easier to reject a hypothesis than to accept it. Our inferences will generally be not of the type "the hypothesis H is true," or even "the hypothesis H is probably true," but of the type "hypotheses A, B and C are probably untrue, but we see no reason to doubt hypothesis H."

For example, suppose we take a sample of a thousand men from the population of the United Kingdom and find their average height to be 5 ft. 8 in. What can we say about the average height of the population as a whole ? We cannot give it with any certainty. We cannot even say, with certainty, that it lies within, say, one inch of 5 ft. 8 in. What we can say, assuming that the sampling technique is sound, will be something to the effect that a hypothesis which supposes that the mean of the whole population is greater than 5 ft. 9 in. or less than 5 ft. 7 in. is *probably* incorrect, but that the data are consistent with the supposition that the mean lies between those limits.

0.41. The theory of sampling is thus closely bound up with the theory of probability. The many problems which arise in this connection are among the most interesting and at times the most difficult which science and philosophy can offer. It is only fair to warn the student that there still exists an important difference of opinion among scientific men about the validity of certain types of statistical inference. In this book we have, so far as we could, avoided these contentious matters, but the advanced student will have to be prepared to face them sooner or later.

The Popular Attitude towards Statistics.

0.42. Finally, to conclude this introduction we may, perhaps, refer to the popular mistrust of statistics and statistical methods.

The layman's attitude towards statistics is admirably summed up in the remark that mankind is divided into two parts, those who say that figures can prove anything and those who assert that they can prove nothing. It must be admitted that this attitude is not unreasonable. From the advertisement hoarding, from the electioneering platform, from the partisan press and from a dozen other sources the man in the street is bombarded with tendentious figures put forward to support some *ex parte* statement. Sometimes such figures are justifiably used to form a basis for the arguments which are built upon them ; more often they give a specious picture of the truth, which may be due to ignorance or inadvertence, but has also been known to be occasioned by a deliberate wish to mislead. The layman is well aware of this fact. His attitude in distrusting all arguments based on figures is that of a reasonable man, who has not the training to distinguish for himself the true from the false, and is therefore inclined to suspect everything.

0.43. We are not concerned here with the vindication of statistics in the public view. We have alluded to the matter in order to remind the student that statistical methods are most dangerous tools in the hands of the inexpert. Few subjects have a wider application ; no subject requires such care in that application. Statistics is one of those sciences whose adepts must exercise the self-restraint of an artist.

THEORY OF ATTRIBUTES—NOTATION AND TERMINOLOGY.

Attributes and Variables.

1.1. The methods of statistics, as defined in the Introduction, deal with quantitative data alone. The quantitative character may, however, arise in two different ways.

In the first place, the observer may note only the *presence* or *absence* of some attribute in a series of objects or individuals, and count how many do or do not possess it. Thus, in a given population, we may count the number of the blind and seeing, the dumb and speaking, or the insane and sane. The quantitative character, in such cases, arises solely in the counting.

In the second place, the observer may note or measure the actual magnitude of some variable character for each of the objects or individuals observed. He may record, for instance, the ages of persons at death, the prices of different samples of a commodity, the statures of men, the numbers of petals in flowers. The observations in these cases are quantitative *ab initio*.

1.2. The methods applicable to the former kind of observations, which may be termed **statistics of attributes,** are also applicable to the latter, or **statistics of variables.** A record of statures of men, for example, may be treated by simply counting all measurements as *tall* that exceed a certain limit, neglecting the magnitude of any excess, and stating the numbers of *tall* and *short* (or more strictly not-tall) on the basis of this classification. Similarly, the methods that are specially adapted to the treatment of statistics of variables, making use of each value recorded, are available to a greater extent than might at first sight seem possible for dealing with statistics of attributes. For example, we may treat the presence or absence of the attribute as corresponding to the changes of a variable which can only possess two values, say 0 and 1. Or, we may assume that we have really to do with a variable character which has been crudely classified, as suggested above, and we may be able, by auxiliary hypotheses as to the nature of this variable, to draw further conclusions. But the methods and principles developed for the case in which the observer only notes the presence or absence of attributes are the simplest and most fundamental, and are best considered first. This and the next four chapters are accordingly devoted to the Theory of Attributes.

Classification with reference to Attributes.

1.3. The objects or individuals that possess the attribute, and those that do not possess it, may be said to be members of two distinct **classes,**

the observer **classifying** the objects or individuals observed. In the simplest case, where attention is paid to one attribute alone, only two mutually exclusive classes are formed. If several attributes are noted, the process of classification may, however, be continued indefinitely. Those that do and do not possess the first attribute may be reclassified according as they do or do not possess the second, the members of each of the sub-classes so formed according as they do or do not possess the third, and so on, every class being divided into two at each step. Thus the members of the population of any district may be classified into males and females ; the members of each sex into sane and insane ; the insane males, sane males, insane females and sane females into blind and seeing. If we were dealing with a number of peas (*Pisum sativum*) of different varieties, they might be classified as tall or dwarf, with green seeds or yellow seeds, with wrinkled seeds or round seeds, so that we should have eight classes— tall with round green seeds, tall with round yellow seeds, tall with wrinkled green seeds, tall with wrinkled yellow seeds, and four similar classes of dwarf plants.

1.4. It may be noticed that the fact of classification does not necessarily imply the existence of either a natural or a clearly defined boundary between the two classes. The boundary may be wholly arbitrary, *e.g.* where prices are classified as above or below some special value, barometer readings as above or below some particular height. The division may also be vague and uncertain : sanity and insanity, sight and blindness, pass into each other by such fine gradations that judgments may differ as to the class in which a given individual should be entered. The possibility of uncertainties of this kind should always be borne in mind in considering statistics of attributes : whatever the nature of the classification, however, natural or artificial, definite or uncertain, the final judgment must be decisive ; any one object or individual must be held either to possess the given attribute or not.

Dichotomy.

1.5. A classification of the simple kind considered, in which each class is divided into two sub-classes and no more, has been termed by logicians **classification,** or, to use the more strictly applicable term, **division by dichotomy** (cutting in two). The classifications of most statistics are not dichotomous, for most usually a class is divided into more than two sub-classes, but dichotomy is the fundamental case. In Chapter 5 the relation of dichotomy to more elaborate (**manifold,** instead of twofold or dichotomous) processes of classification, and the methods applicable to some such cases, are dealt with briefly.

1.6. For theoretical purposes it is necessary to have some simple notation for the classes formed, and for the numbers of observations assigned to each.

The capitals $A, B, C, \ldots$ will be used to denote the several attributes. An object or individual possessing the attribute A will be termed simply A. The class, all the members of which possess the attribute A, will be termed *the class A*. It is convenient to use single symbols also to denote the *absence* of the attributes $A, B, C, \ldots$ We shall employ the Greek letters $a, \beta, \gamma, \ldots$ Thus if A represents the attribute *blindness*, a represents *sight, i.e.* non-blindness ; if B stands for *deafness,* β stands

for *hearing*. Generally " a " is equivalent to " not-A," or *an object or individual not possessing the attribute* A ; *the class* a is equivalent to *the class none of the members of which possesses the attribute* A.

1.7. Combinations of attributes will be represented by juxtapositions of letters. Thus if, as above, A represents *blindness*, B *deafness*, AB represents the combination *blindness and deafness*. If the presence and absence of these attributes be noted, the four classes so formed, viz. AB, $A\beta$, aB, $a\beta$, include respectively the *blind and deaf*, the *blind but not-deaf*, the *deaf but not-blind*, and the *neither blind nor deaf*. If a third attribute be noted, *e.g.* insanity, denoted say by C, the class ABC includes those who are at once *deaf, blind and insane*, $AB\gamma$ those who are *deaf and blind but not insane*, and so on.

Any letter or combination of letters like A, AB, aB, $AB\gamma$, by means of which we specify the characters of the members of a class, may be termed a **class symbol.**

Class-frequencies.

1.8. The number of observations assigned to any class is termed, for brevity, the **frequency** of the class, or the **class-frequency.** Class-frequencies will be denoted by enclosing the corresponding class-symbols in brackets. Thus :

(A)	denotes number of A's,		*i.e.* objects possessing attribute A
(a)	,,	,,	a's,	,, not ,, ,, A
(AB)	,,	,,	AB's,	,, possessing attributes A and B
(aB)	,,	,,	aB's,	,, ,, attribute B but not A
(ABC)	,,	,,	ABC's,	,, ,, attributes A, B and C
(aBC)	,,	,,	aBC's,	,, ,, ,, B and C but not A
($a\beta C$)	,,	,,	$a\beta C$'s,	,, ,, attribute C but neither A nor B

and so on for any number of attributes. If A represent, as in the illustration above, blindness, B deafness, C insanity, the symbols given stand for the numbers of the *blind*, the *not-blind*, the *blind and deaf*, the *deaf but not-blind*, the *blind, deaf and insane*, the *deaf and insane but not-blind*, and the *insane but neither blind nor deaf*, respectively.

Positive Attributes.

1.9. The attributes denoted by capitals ABC . . . may be termed **positive** attributes, and their *contraries* denoted by Greek letters **negative** attributes. If a class-symbol include only capital letters, the class may be termed a positive class ; if only Greek letters, a negative class. Thus the classes A, AB, ABC are positive classes; the classes a, $a\beta$, $a\beta\gamma$, negative classes.

If two classes are such that every attribute in the symbol for the one is the negative or contrary of the corresponding attribute in the symbol for the other, they may be termed **contrary classes** and their frequencies **contrary frequencies** ; *e.g.* AB and $a\beta$, $A\beta$ and aB, $A\beta C$ and $aB\gamma$, are pairs of **contraries.**

1.10. If we make a certain dichotomy with regard to a definite attribute A—such as male sex, blindness or blue eyes—it may be of practical importance to note a possible distinction in the nature of the class not-A. The complementary class may, in fact, either be equally

definite—female sex, ability to see—or it may be a mere heterogeneous remainder, as in our last instance—not-blue-eyed, the not-blue-eyed being brown-eyed, grey-eyed, or even possessing no eyes at all.

Logically, this distinction is difficult to maintain, but practically it is of some importance. The statistical data in official returns are almost always classified according to positive and clearly defined attributes. For example, we are given the numbers of persons dying from typhoid, not the numbers who did *not* die of typhoid ; the number of acres under grass, not the number of acres *not* under grass.

Order of Classes and Class-frequencies.

1.11. The classes obtained by noting, say, n attributes fall into natural groups according to the numbers of attributes used to specify the respective classes, and these natural groups should be borne in mind in tabulating the class-frequencies. A class specified by r attributes may be spoken of as a **class of the rth order** and its frequency as a frequency of the rth order. Thus AB, AC, BC are classes of the second order ; (A), $(A\beta)$, (aBC), $(AB\gamma D)$, class-frequencies of the first, second, third and fourth orders respectively.

Aggregates.

1.12. The classes of one and the same order fall into further groups according to the actual attributes specified. Thus if three attributes A, B, C have been noted, the classes of the second order may be specified by any one of the pairs of attributes AB, AC or BC (and their contraries). The series of classes or class-frequencies given by any one positive class and the classes whose symbols are derived therefrom by substituting Greek letters for one or more of the italic capital letters in every possible way will be termed an **aggregate**. Thus (AB), $(A\beta)$, (aB), $(a\beta)$ form an aggregate of frequencies of the second order, and the twelve classes of the second order which can be formed where three attributes have been noted may be grouped into three such **aggregates**.

1.13. Class-frequencies should, in tabulating, be arranged so that frequencies of the same order and frequencies belonging to the same aggregate are kept together. Thus the frequencies for the case of three attributes should be grouped as given below, the whole number of observations denoted by the letter N being reckoned as a frequency of order zero, since no attributes are specified.

$$
\begin{array}{llll}
\text{Order 0.} & N \\
\text{Order 1.} & (A) & (B) & (C) \\
 & (a) & (\beta) & (\gamma) \\
\text{Order 2.} & (AB) & (AC) & (BC) \\
 & (A\beta) & (A\gamma) & (B\gamma) \\
 & (aB) & (aC) & (\beta C) \\
 & (a\beta) & (a\gamma) & (\beta\gamma) \\
\text{Order 3.} & (ABC) & (aBC) \\
 & (AB\gamma) & (aB\gamma) \\
 & (A\beta C) & (a\beta C) \\
 & (A\beta\gamma) & (a\beta\gamma) \\
\end{array}
\qquad (1.1)
$$

The Total Number of Class-frequencies.

1.14. In such a complete table for the case of three attributes, twenty-seven distinct frequencies are given: 1 of order zero, 6 of the first order, 12 of the second and 8 of the third.

In general, for n attributes, there are 3^n distinct class-frequencies, if we count N as a frequency of order 0.

To demonstrate this, let us consider the number of classes of different orders.

Of order 0 there is one class N.

Of order 1 there are $2n$ classes, for classes of this order contain only one symbol, and each of the n attributes contributes two symbols, one of the type A and one of the type a.

Of order 2 there are $\dfrac{n(n-1)}{2} \times 2^2$ classes, for each class contains two symbols, two attributes can be chosen from n in $\dfrac{n(n-1)}{2}$ ways, and each pair gives rise to 2^2 different frequencies of the types (AB), $(A\beta)$, (aB) and $(a\beta)$.

Similarly, it may be seen that of order r there are

$$\frac{n(n-1) \ . \ . \ . \ (n-r+1)}{r!} \times 2^r$$

classes.

Hence, the total number of class-frequencies is

$$1 + n \cdot 2 + \frac{n(n-1)}{2} \cdot 2^2 + \ . \ . \ . \ + \frac{n(n-1) \ . \ . \ . \ (n-r+1)}{r!} \times 2^r + \ . \ . \ .$$

and this is the binomial expansion of $(1+2)^n = 3^n$.

It is clear that if n is at all large the number of class-frequencies will be very great. For instance, if $n = 6$, the number is 729.

1.15. Fortunately, however, the class-frequencies are not independent of one another, and it is not necessary, in order to specify the data completely, to give every class-frequency.

In the first place, let us note the simple result *that any class-frequency can always be expressed in terms of class-frequencies of higher order.* For the whole number of observations must clearly be equal to the number of A's added to the number of a's, *i.e.*

$$N = (A) + (a) \qquad . \qquad . \qquad . \qquad . \quad (1.2)$$

Similarly, the number of A's is equal to the number of A's which are B's added to the number of A's which are β's, *i.e.*

$$(A) = (AB) + (A\beta) \qquad . \qquad . \qquad . \qquad . \quad (1.3)$$

Similarly,

$$(AB) = (ABC) + (AB\gamma) \qquad . \qquad . \qquad . \quad (1.4)$$

and so on.

Ultimate Class-frequencies.

1.16. It follows at once from the result we have just given that every class-frequency can be expressed in terms of the frequencies of the highest

order, *i.e.* of order n. For any frequency can be analysed into higher frequencies, and the process need only stop when we have reached the frequencies of highest order. For example, with three attributes,

$$(A) = (AB) + (A\beta)$$
$$= (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma)$$

The classes specified by n attributes, *i.e.* those of the highest order, are termed the **ultimate class-frequencies.**

Our result may then be expressed in the form : *Every class-frequency can be expressed as the sum of certain of the ultimate class-frequencies.* To specify the data completely it is, therefore, only necessary to give the ultimate class-frequencies.

Example 1.1.—(See ref. (69).) A number of school-children were examined for the presence or absence of certain defects of which three chief descriptions were noted : A, development defects ; B, nerve signs ; C, low nutrition.

Given the following ultimate frequencies, find the frequencies of the positive classes, including the whole number of observations N :—

(ABC)	57	(aBC)	78
$(AB\gamma)$	281	$(aB\gamma)$	670
$(A\beta C)$	86	$(a\beta C)$	65
$(A\beta\gamma)$	453	$(a\beta\gamma)$	8310

The whole number of observations N is equal to the grand total : $N = 10,000$.

The frequency of any first-order class, *e.g.* (A), is given by the total of the four third-order frequencies the class-symbols for which contain the same letter :

$$(ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma) = (A) = 877$$

Similarly, the frequency of any second-order class, *e.g.* (AB), is given by the total of the two third-order frequencies the class-symbols for which both contain the same pair of letters :

$$(ABC) + (AB\gamma) = (AB) = 338$$

The complete results are :

N	10,000	(AB)	338
(A)	877	(AC)	143
(B)	1,086	(BC)	135
(C)	286	(ABC)	57

The Number of Ultimate Class-frequencies.

1.17. The class-frequencies of highest order each contain n symbols. Now each letter corresponding to a particular attribute may be written in two ways: A or a, B or β, etc. Hence the total number of possible symbols is

$$2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times \ldots = 2^n$$

and this is the number of ultimate class-frequencies.

Hence the 3^n frequencies may all be expressed in terms of the 2^n ultimate frequencies. For example, if $n = 6$, the 729 frequencies can be

written in terms of 64 ultimate class-frequencies, which specify the data completely.

Fundamental Sets.

1.18. The ultimate frequencies are, however, not the only set which specify the whole of the data. In fact, any set will serve the purpose provided that (a) they are 2^n in number, and (b) they are algebraically independent; that is to say, when they are written symbolically no one can be expressed in terms of some or all of the others.

We may call such a set of frequencies a **fundamental set.**

The Positive Class-frequencies form a Fundamental Set.

1.19. The **positive** class-frequencies, including under this head the total number of observations N, form one such set. They are algebraically independent; no one positive class-frequency can be expressed wholly in terms of the others. Their number is, moreover, 2^n, as may be readily seen from the fact that if the Greek letters are struck out of the symbols for the ultimate classes, they become the symbols for the positive classes, with the exception of $\alpha\beta\gamma \ldots$ for which N must be substituted. Alternatively we may, in the manner of 1.14, prove the result by considering the number of positive class-frequencies of each order. The number is made up as follows :—

Order 0. (The whole number of observations) . . . 1

Order 1. (The number of attributes noted) . . . . n

Order 2. (The number of combinations of n things 2 together) $\dfrac{n(n-1)}{1 \cdot 2}$

Order 3. (The number of combinations of n things 3 together)
$$\dfrac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3}$$

and so on. But the series

$$1 + n + \frac{n(n-1)}{1 \cdot 2} + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} + \ldots$$

is the binomial expansion of $(1+1)^n$ or 2^n; therefore the total number of positive classes is 2^n.

1.20. The set of positive class-frequencies is a most convenient one for both theoretical and practical purposes.

Compare, for instance, the two forms of statement, in terms of the ultimate and the positive classes respectively, as given in Example 1.1. The latter gives directly the whole number of observations and the totals of A's, B's and C's. The former gives none of these fundamentally important figures without the performance of more or less lengthy additions. Further, the latter gives the second-order frequencies (AB), (AC) and (BC), which are necessary for discussing the relations subsisting between A, B and C, but are only indirectly given by the frequencies of the ultimate classes.

1.21. We are now able to indicate the applications of the foregoing analysis to some practical problems.

2

The typical problem which arises in this connection is the following : Given certain class-frequencies, to find them all.

In the first place, we may remark at once that unless 2^n independent class-frequencies are given the problem is insoluble. We might be able to find some of the frequencies, but it is certain that we could not find every one. We shall reserve to a later chapter the consideration of what can be done with such incomplete data. In the examples of this chapter we shall deal only with data which specify the problem completely.

Example 1.2.—Given the positive class-frequencies of Example 1.1, to find all the class-frequencies.

The data are :

$N = 10,000$; $(A) = 877$; $(B) = 1086$; $(C) = 286$; $(AB) = 338$; $(AC) = 143$; $(BC) = 135$; $(ABC) = 57$.   ✱

We have :

$$(AB) = (AB\gamma) + (ABC)$$

or

$$338 = (AB\gamma) + 57$$

i.e.

$$(AB\gamma) = 281$$

Similarly, from (AC) and (BC) we find :

$$(A\beta C) = 86$$
$$(aBC) = 78$$

This gives us the three ultimate class-frequencies which contain only one Greek letter. For the others,

$$\begin{aligned}(a\beta C) &= (\beta C) - (A\beta C)\\ &= (C) - (BC) - (A\beta C)\\ &= 286 - 135 - 86\\ &= 65\end{aligned}$$

Similarly, we have :

$$(A\beta\gamma) = 453$$
$$(aB\gamma) = 670$$

Finally,

$$\begin{aligned}(a\beta\gamma) &= (\beta\gamma) - (A\beta\gamma)\\ &= (\gamma) - (B\gamma) - (A\beta\gamma)\\ &= N - (C) - \{(B) - (BC)\} - (A\beta\gamma)\\ &= 10,000 - 286 - 951 - 453\\ &= 8310\end{aligned}$$

We can now calculate any class-frequency by expressing it in terms of the ultimate class-frequencies, *e.g.*

$$\begin{aligned}(a\gamma) &= (aB\gamma) + (a\beta\gamma)\\ &= 670 + 8310\\ &= 8980\end{aligned}$$

It is, of course, also possible to calculate these frequencies by expressing them directly in terms of the given frequencies, *e.g.*

$$(a\gamma) = (\gamma) - (A\gamma)$$
$$= N - (C) - \{(A) - (AC)\}$$
$$= 10,000 - 286 - 877 + 143$$
$$= 8980$$

Example 1.3.—In a free vote in the House of Commons, 600 members voted. 300 Government members representing English constituencies (including Welsh) voted in favour of the motion. 25 Opposition members representing Scottish constituencies voted against the motion. The Government majority among those who voted was 96. 135 of the members voting represented Scottish constituencies. 18 Government members voted against the motion. 102 Scottish members voted in favour of the motion. The motion was carried by 310 votes. Analyse the voting according to the nationality of the constituencies and party.

Denoting the Government and Opposition parties by A and a respectively, voting for and against the motion by B and β, and English and Scottish members by C and γ respectively, our data, in the order of the question, are :

$$
\begin{array}{lll}
N = 600 & \qquad . \qquad . \qquad . \qquad . & (a) \\
(ABC) = 300 & \qquad . \qquad . \qquad . \qquad . & (b) \\
(a\beta\gamma) = 25 & \qquad . \qquad . \qquad . \qquad . & (c) \\
(A) - (a) = 96 & \qquad . \qquad . \qquad . \qquad . & (d) \\
(\gamma) = 135 & \qquad . \qquad . \qquad . \qquad . & (e) \\
(A\beta) = 18 & \qquad . \qquad . \qquad . \qquad . & (f) \\
(B\gamma) = 102 & \qquad . \qquad . \qquad . \qquad . & (g) \\
(B) - (\beta) = 310 & \qquad . \qquad . \qquad . \qquad . & (h)
\end{array}
$$

We wish to find the ultimate class-frequencies.

Let us note first of all that there are $2^3 = 8$ equations here. We therefore expect them to give us the eight ultimate classes. Equations (b) and (c) already give us two.

From (a) we have :
$$N = (A) + (a) = 600$$
From (d) :
$$(A) - (a) = 96$$
Hence,
$$
\begin{array}{lll}
(A) = 348 & \qquad . \qquad . \qquad . \qquad . & (i) \\
(a) = 252 & \qquad . \qquad . \qquad . \qquad . & (j)
\end{array}
$$

Similarly, from (a) and (h) we obtain :
$$
\begin{array}{lll}
(B) = 455 & \qquad . \qquad . \qquad . \qquad . & (k) \\
(\beta) = 145 & \qquad . \qquad . \qquad . \qquad . & (l)
\end{array}
$$

From (a) and (e) we have :
$$(C) = N - (\gamma) = 465 \qquad . \qquad . \qquad . \qquad (m)$$

We have thus found all first-order frequencies.

(i) and (f) give
$$(AB) = (A) - (A\beta)$$
$$= 330 \qquad . \qquad . \qquad . \qquad . \qquad (n)$$

(k) and (g) give

$$(BC) = (B) - (B\gamma)$$
$$= 353 \qquad . \qquad . \qquad . \qquad . \qquad (o)$$

We also have:

$$(\alpha\beta\gamma) = (\beta\gamma) - (A\beta\gamma)$$
$$= (\gamma) - (B\gamma) - \{(A) - (AC) - (AB) + (ABC)\}$$

and substituting the known values on the right and the value of ($\alpha\beta\gamma$), we have

$$25 = 135 - 102 - 348 + (AC) + 330 - 300$$
$$(AC) = 310 \qquad . \qquad . \qquad . \qquad . \qquad (p)$$

From (n) and (b) we get

$$(AB\gamma) = (AB) - (ABC) = 30 \qquad . \qquad . \qquad . \qquad (q)$$

From (o) and (b) we get, similarly,

$$(\alpha BC) = 53 \qquad . \qquad . \qquad . \qquad . \qquad (r)$$

From (p) and (b) we get

$$(A\beta C) = 10 \qquad . \qquad . \qquad . \qquad . \qquad (s)$$

From (e) and (g):

$$(\beta\gamma) = (\gamma) - (B\gamma) = 33$$

Hence,

$$(A\beta\gamma) = (\beta\gamma) - (\alpha\beta\gamma) = 8 \qquad . \qquad . \qquad . \qquad (t)$$

From (f) and (l):

$$(\alpha\beta) = 127$$

Hence,

$$(\alpha\beta C) = (\alpha\beta) - (\alpha\beta\gamma)$$
$$= 102 \qquad . \qquad . \qquad . \qquad . \qquad (u)$$

Finally, $N = $ sum of ultimate class-frequencies, and this gives

$$(\alpha B\gamma) = 72 \qquad . \qquad . \qquad . \qquad . \qquad (v)$$

This straightforward but rather heavy analysis has therefore given us the eight ultimate class-frequencies in equations (b), (c), (q), (r), (s), (t), (u) and (v).

1.22. The data encountered in practice are rarely dichotomised according to more than three or four variables, and the student should experience little difficulty in expressing any class-frequency in terms of the known class-frequencies, either directly, or by first finding the ultimate class-frequencies and then expressing the desired frequency in terms of them.

It is, however, interesting to note the general result that the class symbols can be treated as operators and multiplied together like algebraical quantities. Let us write $A \cdot N$ for the operation of dichotomising N according to A, and write

$$A \cdot N = (A)$$

which is the symbolic way of saying that if we dichotomise N according to A we get a class-frequency equal to (A). We can similarly put

$$a \cdot N = (a)$$

Adding these two, and putting $A \cdot N + a \cdot N$ equal to $(A+a) \cdot N$, we have :

$$(A + a) \cdot N = N$$

so that we may take

$$A + a = 1$$

In any symbolic expression we can therefore replace the operators A or a by $1 - a$, $1 - A$, respectively.

Furthermore, since $(AB) = A \cdot (B) = B \cdot (A)$, we may take the symbol $AB \cdot N$ to be the dichotomy of N according to both A and B, and equate it to (AB). A little reflection will show that the operative symbols therefore obey the ordinary laws of algebra and in particular may be multiplied together.

For example, we have :

$$\begin{aligned}
(a\beta) = a\beta \cdot N &= (1 - A)(1 - B) \cdot N \\
&= (1 - A - B + AB) \cdot N \\
&= N - (A) - (B) + (AB) .
\end{aligned}$$ (1.5)

And, similarly,

$$\begin{aligned}
(a\beta\gamma) = a\beta\gamma \cdot N \\
&= (1 - A)(1 - B)(1 - C) \cdot N \\
&= (1 - A - B - C + AB + BC + AC - ABC) \cdot N \\
&= N - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC) .
\end{aligned}$$ (1.6)

Similar results could, of course, be obtained by step-by-step substitution ; for instance,

$$\begin{aligned}
(a\beta) &= (a) - (aB) \\
&= N - (A) - (B) + (AB)
\end{aligned}$$

1.23. The symbolism we have discussed in this chapter is also of use in deducing results of a less definite character expressible by inequalities.

Example 1.4.—In a war between White and Red forces there are more Red soldiers than White ; there are more armed Whites than unarmed Reds ; there are fewer armed Reds with ammunition than unarmed Whites without ammunition. Show that there are more armed Reds without ammunition than unarmed Whites with ammunition.

Writing A to denote the property of being a White soldier, and hence a to denote the property of being a Red soldier ; writing B and β to denote armed and unarmed, respectively ; and writing C and γ to denote the possession or non-possession of ammunition, respectively, our data are :

$$(a) > (A) .$$ (a)
$$(AB) > (a\beta)$$ (b)
$$(A\beta\gamma) > (aBC)$$ (c)

We have to show that

$$(aB\gamma) > (A\beta C)$$

From (a), considering the dichotomy of each side according to B, we have :

$$(aB) + (a\beta) > (AB) + (A\beta)$$

Substituting for (AB) from (b) in this inequality,

$$(aB) + (a\beta) > (a\beta) + (A\beta)$$

and hence,

$$(aB) > (A\beta) \qquad . \qquad . \qquad . \qquad . \qquad (d)$$

From this, considering the dichotomy of each side according to C, we have :

$$(aBC) + (aB\gamma) > (A\beta C) + (A\beta \gamma)$$

and in virtue of (c) this gives

$$(aB\gamma) > (A\beta C)$$

which is the required result.

1.24. The symbols of our notation are, it should be remarked, used in an inclusive sense, the symbol A, for example, signifying an object or individual possessing the attribute A with or without others. This seems to be the only natural use of the symbol, but at least one notation has been constructed on an *exclusive* basis, the symbol A denoting that the object or individual possesses the attribute A, but not B or C or D, or whatever other attributes have been noted. An exclusive notation is apt to be relatively cumbrous and also ambiguous, for the reader cannot know what attributes a given symbol excludes until he has seen the whole list of attributes of which note has been taken, and this list he must bear in mind. The statement that the symbol A is used exclusively cannot mean, obviously, that the object referred to possesses only the attribute A and no others whatever ; it merely excludes the other attributes noted in the particular investigation. Adjectives, as well as the symbols which may represent them, are naturally used in an inclusive sense, and care should therefore be taken, when classes are verbally described, that the description is complete, and states what, if anything, is excluded as well as what is included, in the same way as our notation. The terminology of some tables in our older English census has not, in this respect, been quite clear. The " Blind " includes those who are " Blind and Dumb," or " Blind, Dumb and Lunatic," and so forth. But the heading " Blind and Dumb," in the table relating to " combined infirmities," is used in the sense " Blind and Dumb, but not Lunatic or Imbecile," etc., and so on for the others. In the first table the headings are inclusive, in the second exclusive.

SUMMARY.

1. A collection of individuals may be divided into two classes according to whether they do or do not possess a particular attribute. This process is called dichotomy.

2. Continued dichotomy according to n attributes gives rise to 3^n classes.

3. The frequencies in these classes can be expressed in terms of the 2^n ultimate class-frequencies, or of the 2^n positive class-frequencies.

4. Given 2^n independent class-frequencies, all the class-frequencies may be calculated by simple arithmetical processes.

EXERCISES.

1.1. (Figures from ref. (69).)   The following are the numbers of boys observed with certain classes of defects amongst a number of school-children.   *A* denotes development defects; *B*, nerve signs; *C*, low nutrition.

(*ABC*)	149	(*aBC*)	204
(*ABγ*)	738	(*aBγ*)	1,762
(*AβC*)	225	(*aβC*)	171
(*Aβγ*)	1,196	(*aβγ*)	21,842

Find the frequencies of the *positive* classes.

1.2. (Figures from ref. (69).)   The following are the frequencies of the positive classes for the girls in the same investigation:—

N	23,713	(*AB*)	587
(*A*)	1,618	(*AC*)	428
(*B*)	2,015	(*BC*)	335
(*C*)	770	(*ABC*)	156

Find the frequencies of the ultimate classes.

1.3. (Figures from *Census, England and Wales, 1891*, vol. 3.)   Convert the census statement as below into a statement in terms of (*a*) the positive, (*b*) the ultimate class-frequencies.   *A* =blindness, *B* =deaf-mutism, *C* =mental derangement.

N	29,002,525	(*ABγ*)	82
(*A*)	23,467	(*AβC*)	380
(*B*)	14,192	(*aBC*)	500
(*C*)	97,383	(*ABC*)	25

1.4. (*Cf.* Mill's " *Logic*," bk. 3, ch. 17, and ref. (65).)   Show that if *A* occurs in a larger proportion of the cases where *B* is than where *B* is not, then *B* will occur in a larger proportion of the cases where *A* is than where *A* is not: *i.e.* given $(AB)/(B) > (A\beta)/(\beta)$, show that $(AB)/(A) > (aB)/(a)$.

1.5. (*Cf.* De Morgan, " *Formal Logic*," p. 163, and ref. (65).)   Most *B*'s are *A*'s, most *B*'s are *C*'s: find the least number of *A*'s that are *C*'s, *i.e.* the lowest possible value of (*AC*).

· 1.6. Given that

$$(A) = (a) = (B) = (\beta) = \tfrac{1}{2}N$$

show that

$$(AB) = (a\beta), \quad (A\beta) = (aB)$$

1.7. (*Cf.* ref. (78), Section 9, "Case of equality of contraries.")   Given that

$$(A) = (a) = (B) = (\beta) = (C) = (\gamma) = \tfrac{1}{2}N$$

and also that

$$(ABC) = (a\beta\gamma)$$

show that

$$2(ABC) = (AB) + (AC) + (BC) - \tfrac{1}{2}N$$

1.8. Measurements are made on a thousand husbands and a thousand wives. If the measurements of the husbands exceed the measurements of the wives in 800 cases for one measurement, in 700 cases for another, and in 660 cases for both measurements, in how many cases will both measurements on the wife exceed the measurements on the husband ?

1.9. 100 children took three examinations.   40 passed the first, 39 passed the second and 48 passed the third.   10 passed all three, 21 failed all three, 9 passed the first two and failed the third, 19 failed the first two and passed the third.   Find how many children passed at least two examinations.

Show that for the question asked certain of the given frequencies are not necessary. Which are they?

Show further that the data are not sufficient to permit of the determination of the ultimate class-frequencies.

1.10. (Lewis Carroll, "*A Tangled Tale*," 1881.) In a very hotly fought battle 70 per cent. at least of the combatants lost an eye, 75 per cent. at least lost an ear, 80 per cent. at least lost an arm and 85 per cent. at least lost a leg. How many at least must have lost all four?

1.11. Show that for n attributes $A, B, C, \ldots M$,

$$(ABC \ldots M) \geqslant \{(A) + (B) + (C) + \ldots + (M)\} - (n-1)N$$

where N is the total frequency; and hence generalise the result of Exercise 1.10.

CHAPTER 2.

CONSISTENCE OF DATA.

Universe of Discourse.

2.1. Any statistical inquiry is necessarily confined to a certain time, space or material. An investigation on the prevalence of unemployment, for instance, may be limited to England, to England in 1931, to English males in 1931, or even to English males over 50 years of age in 1931, and so on.

For actual work on any given subject, no term is required to denote the material to which the work is so confined : the limits are specified, and that is sufficient. But for theoretical purposes some term is almost essential to avoid circumlocution. The expression the **universe of discourse,** or simply the **universe,** used in this sense by writers on logic, may be adopted as familiar and convenient.

2.2. The **universe,** like any **class,** may be considered as specified by an enumeration of the attributes common to all its members; *e.g.* taking the illustration of **2.1,** those attributes implied by the predicates *English, male, over 50 years of age, living in 1931.* It is not, in general, necessary to introduce a special letter into the class-symbols to denote the attributes common to all members of the universe. We know that such attributes must exist, and the common symbol can be understood.

In strictness, however, the symbol ought to be written : if, say, U denote the combination of attributes, English —male—over 50— living in 1931, A unemployed, B married, we should strictly use the symbols :

(U) = Number of English males over 50 living in 1931
(UA) =   „   unemployed English males over 50 living in 1931
(UB) =   „   married   „   „   „
(UAB) =   „   unemployed and married English males over 50 living in 1931

instead of the simpler symbols N, (A), (B), (AB). Similarly, the general relations of equations (1.2), (1.3) and (1.4), page 15, using U to denote the common attributes of all the members of the universe and (U) consequently the total number of observations N, should in strictness be written in the form :

$$(U) = (UA) + (Ua) = (UB) + (U\beta) = \text{etc.}$$
$$= (UAB) + (UA\beta) + (UaB) + (Ua\beta) = \text{etc.}$$
$$(UA) = (UAB) + (UA\beta) = (UAC) + UA\gamma) = \text{etc.}$$
$$(UAB) = (UABC) + (UAB\gamma) = \text{etc.}$$

Specifying the Universe.

2.3. Clearly, however, we might have used any other symbol instead of U to denote the attributes common to all the members of the universe, *e.g.* A or B or AB or ABC, writing in the latter case:

$$(ABC) = (ABCD) + (ABC\delta)$$

and so on. Hence *any attribute or combination of attributes common to all the class-symbols in an equation may be regarded as specifying the universe within which the equation holds good.* Thus the equation just written may be read in words : " The number of objects or individuals in the universe ABC is equal to the number of D's together with the number of not-D's within the same universe." The equation

$$(AC) = (ABC) + (A\beta C)$$

may be read : " The number of A's is equal to the number of A's that are B's together with the number of A's that are not-B's *within the universe C.*"

2.4. The more complex relations between class-frequencies may be derived from the simpler ones very readily by the process of **specifying the universe.** Thus, starting from the simple equation

$$(a) = N - (A)$$

we have, by specifying the universe as β,

$$(a\beta) = (\beta) - (A\beta)$$
$$= N - (A) - (B) + (AB)$$

Specifying the universe, again, as γ, we have :

$$(a\beta\gamma) = (\gamma) - (A\gamma) - (B\gamma) + (AB\gamma)$$
$$= N - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC)$$

Consistence.

2.5. Any class-frequencies which have been or might have been observed within one and the same universe may be said to be **consistent** with one another. They conform with one another, and do not in any way conflict.

The conditions of consistence are some of them simple, but others are by no means of an intuitive character. Suppose, for instance, the following data are given :—

N	1000	(AB)	42
(A)	525	(AC)	147
(B)	312	(BC)	86
(C)	470	(ABC)	25

—there is nothing obviously wrong with the figures. Yet they are certainly inconsistent. They might have been observed at different times, in different places or on different material, but they cannot have been observed in one and the same universe. They imply, in fact, a negative value for $(a\beta\gamma)$:

$$(a\beta\gamma) = 1000 - 525 - 312 - 470 + 42 + 147 + 86 - 25$$
$$= 1000 - 1307 + 275 - 25$$
$$= -57$$

Clearly no class-frequency can be negative. If the figures, consequently, are alleged to be the result of an actual inquiry in a definite universe, there must have been some miscount or misprint.

Condition for Consistence.

2.6. It is, in fact, the necessary and sufficient condition for the consistence of a set of independent class-frequencies that no ultimate class-frequency be negative. It is necessary for the obvious reason that no class-frequency occurring by counting real attributes can be negative ; it is sufficient because, given any non-negative set of 2^n numbers, we can always imagine a real universe with n dichotomies which should have these numbers for its ultimate class-frequencies, and it is impossible for this real universe to give inconsistent results.

Hence to test the consistence of a set of 2^n algebraically independent class-frequencies we need only calculate the ultimate class-frequencies and ascertain whether any one is negative. If it is, the data are inconsistent. If no ultimate frequency is negative, the data are consistent.

Consistence of Positive Class-frequencies.

2.7. For data given by a heterogeneous collection of class-frequencies, consistence is best tested by actually calculating the ultimate frequencies. We saw in the last chapter, however, that the positive class-frequencies hold a peculiar position in that many data encountered in practice are given entirely in terms of them alone. To save the trouble of calculating the ultimate frequencies from them, we proceed to discuss the form which the consistence conditions assume when expressed entirely in terms of the positive class-frequencies. These conditions may be expressed symbolically by expanding the ultimate in terms of the positive frequencies, and writing each such expansion not less than zero. We will consider the cases of one, two and three attributes in turn.

2.8. If only one attribute be noted, say A, the positive frequencies are N and (A). The ultimate frequencies are (A) and (a), where

$$(a) = N - (A)$$

The conditions of consistence are therefore simply

$$(A) \not< 0 \qquad N - (A) \not< 0$$

or, more conveniently expressed,

$$(a) \quad (A) \not< 0 \qquad (b) \quad (A) \not> N \qquad . \qquad . \qquad . \quad (2.1)$$

These conditions are obvious : the number of A's cannot be less than zero, nor exceed the whole number of observations.

2.9. If two attributes be noted there are four ultimate frequencies (AB), $(A\beta)$, (aB), $(a\beta)$. The following conditions are given by expanding each in terms of the frequencies of positive classes :—

$$
\begin{array}{llll}
(a) & (AB) \not< 0 & \text{or } (AB) \text{ would be negative} \\
(b) & (AB) \not< (A) + (B) - N & \text{,, } (a\beta) \text{ ,, ,,} \\
(c) & (AB) \not> (A) & \text{,, } (A\beta) \text{ ,, ,,} \\
(d) & (AB) \not> (B) & \text{,, } (aB) \text{ ,, ,,}
\end{array}
\qquad (2.2)
$$

(*a*), (*c*) and (*d*) are obvious ; (*b*) is perhaps a little less obvious, and is occasionally forgotten. It is, however, of precisely the same type as the other three. None of these conditions is really of a new form, but may be derived at once from (2.1) (*a*) and (2.1) (*b*) by specifying the universe as *B* or as β respectively. The conditions (2.2) are therefore really covered by (2.1).

2.10. But a further point arises as regards such a system of limits as is given by (2.2). The conditions (*a*) and (*b*) give lower or minor limits to the value of (*AB*) ; (*c*) and (*d*) give upper or major limits. If either major limit be less than either minor limit the conditions are impossible, and it is necessary to see whether (*A*) and (*B*) can take such values that this may be the case.

Expressing the condition that the major limits must be not less than the minor, we have :

$$
\begin{array}{ll}
(A) \not< 0 \ \big\} & (B) \not< 0 \ \big\} \\
(A) \not> N \big\} & (B) \not> N \big\}
\end{array}
$$

These are simply the conditions of the form (2.1). If, therefore, (*A*) and (*B*) fulfil the conditions (2.1), the conditions (2.2) must be possible. The conditions (2.1) and (2.2) therefore give all the conditions of consistence for the case of two attributes, conditions of an extremely simple and obvious kind.

2.11. Now consider the case of three attributes. There are eight ultimate frequencies. Expanding the ultimate in terms of the positive frequencies, and expressing the condition that each expansion is not less than zero, we have :

or the frequency given below will be negative

$$
\begin{array}{lll}
(a) & (ABC) \not< 0 & (ABC) \ \Big\rbrace \\
(b) & \not< (AB)+(AC)-(A) & (A\beta\gamma) \\
(c) & \not< (AB)+(BC)-(B) & (\alpha B\gamma) \\
(d) & \not< (AC)+(BC)-(C) & (\alpha\beta C) \\
(e) & \not> (AB) & (AB\gamma) \quad (2.3) \\
(f) & \not> (AC) & (A\beta C) \\
(g) & \not> (BC) & (\alpha BC) \\
(h) & \not> (AB)+(AC)+(BC)-(A)-(B)-(C)+N & (\alpha\beta\gamma) \ \Big\rbrace
\end{array}
$$

These, again, are not conditions of a new form. We leave it as an exercise for the student to show that they may be derived from (2.1) (*a*) and (2.1) (*b*) by specifying the universe in turn as *BC*, *B*γ, β*C* and $\beta\gamma$. The *two* conditions holding in *four* universes give the eight inequalities above.

2.12. As in the last case, however, these conditions will be impossible to fulfil if any one of the major limits (*e*)–(*h*) be less than any one of the minor limits (*a*)–(*d*). The values on the right must be such as to make no major limit less than a minor.

There are four major and four minor limits, or sixteen comparisons in all to be made. But twelve of these, the student will find, only lead back to conditions of the form (2.2) for (*AB*), (*AC*) and (*BC*) respectively. The four comparisons of expansions due to contrary frequencies ((*a*) and (*h*), (*b*) and (*g*), (*c*) and (*f*), (*d*) and (*e*)) alone lead to new conditions, viz.

$$\left.\begin{array}{ll} (a) & (AB)+(AC)+(BC) \ll (A)+(B)+(C)-N \\ (b) & (AB)+(AC)-(BC) \gg (A) \\ (c) & (AB)-(AC)+(BC) \gg (B) \\ (d) & (AB)+(AC)+(BC) \gg (C) \end{array}\right\} \qquad (2.4)$$

2.13. These are conditions of a wholly new type, not derivable in any way from those given under (2.1) and (2.2). They are conditions for the consistence of the second-order frequencies *with each other*, whilst the inequalities of the form (2.2) are only conditions for the consistence of the second-order frequencies with those of lower orders. Given any two of the second-order frequencies, *e.g.* (AB) and (AC), the conditions (2.4) give limits for the third, viz. (BC).

Incomplete Data.

2.14. We can now take up a question which we set aside in Chapter 1, namely, that of the inferences which may be drawn from data which, though giving us a certain amount of information in the shape of class-frequencies, yet are insufficient to enable us to calculate all the class-frequencies.

The form of the consistence conditions (2.4) shows that a knowledge of certain class-frequencies allows us to assign limits to others, even though we may not be able to find the actual values of those others. The following will serve as illustrations of the statistical uses of the conditions :—

Example 2.1.—Given that $(A)=(B)=(C)=\frac{1}{2}N$ and 80 per cent. of the A's are B's, 75 per cent. of A's are C's, find the limits to the percentage of B's that are C's.

The data are :
$$\frac{2(AB)}{N}=0\cdot8 \qquad \frac{2(AC)}{N}=0\cdot75$$

and the conditions (2.4) give :

$$\begin{array}{lll} (a) & \dfrac{2(BC)}{N} \ll 1 & -0\cdot8 \quad -0\cdot75 \\ (b) & \ll 0\cdot8+0\cdot75-1 \\ (c) & \gg 1 \quad -0\cdot8 \quad +0\cdot75 \\ (d) & \gg 1 \quad +0\cdot8 \quad -0\cdot75 \end{array}$$

(a) gives a negative limit and (d) a limit greater than unity ; hence they may be disregarded. From (b) and (c) we have :

$$\frac{2(BC)}{N} \ll 0\cdot55 \qquad \frac{2(BC)}{N} \gg 0\cdot95$$

—that is to say, not less than 55 per cent. nor more than 95 per cent. of the B's can be C's.

Example 2.2.—If a report gives the following frequencies as actually observed, show that there must be a misprint or mistake of some sort, and that possibly the misprint consists in the dropping of a 1 before the 85 given as the frequency (BC) :—

N 1000

(A)	510	(AB)	189
(B)	490	(AC)	140
(C)	427	(BC)	85

From (2.4) (a) we have :

$$(BC) \nless 510 + 490 + 427 - 1000 - 189 - 140$$
$$\nless 98$$

But 85 < 98, therefore it cannot be the correct value of (BC).
If we read 185 for 85 all the conditions are fulfilled.

Example 2.3.--In a certain set of 1000 observations $(A) = 45$, $(B) = 23$, $(C) = 14$. Show that whatever the percentages of B's that are A's and of C's that are A's, it cannot be inferred that any B's are C's.

The conditions (2.4) (a) and (b) give the lower limit of (BC), which is required. We find :

$$(a) \quad \frac{(BC)}{N} \nless -\frac{(AB)}{N} - \frac{(AC)}{N} - 0.918$$

$$(b) \quad \frac{(BC)}{N} \nless \frac{(AB)}{N} + \frac{(AC)}{N} - 0.045$$

The first limit is clearly negative. The second must also be negative, since $(AB)/N$ cannot exceed 0.023 nor $(AC)/N$ 0.014. Hence we cannot conclude that there is any limit to (BC) greater than 0. This result is indeed immediately obvious when we consider that, even if all the B's were A's, and of the remaining 22 A's 14 were C's, there would still be 8 A's that were neither B's nor C's.

2.15. The student should note the result of the last example, as it illustrates the sort of result at which one may often arrive by applying the conditions (2.4) to practical statistics. For given values of N, (A), (B), (C), (AB) and (AC), it will often happen that *any* value of (BC) not less than zero (or, more generally, not less than either of the lower limits (2.2) (a) and (2.2) (b)) will satisfy the conditions (2.4), and hence no true inference of a lower limit is possible. The argument of the type " So many A's are B's and so many B's are C's that we must expect some A's to be C's " must be used with caution.

2.16. Where the data are not given in terms of the positive or of the ultimate class-frequencies, and cannot readily be thrown into such a form, the device illustrated in the following example is often useful :—

Example 2.4.—Among the adult population of a certain town 50 per cent. of the population are male, 60 per cent. are wage-earners and 50 per cent. are 45 years of age or over. 10 per cent. of the males are not wage-earners and 40 per cent. of the males are under 45. Can we infer anything about what percentage of the population of 45 or over are wage-earners ?

Denoting the attributes male, wage-earner and 45 years old or more by A, B and C, respectively, and letting $N = 100$ for convenience, our data are :

$$(A) = 50$$
$$(B) = 60$$
$$(C) = 50$$
$$(A\beta) = 5$$
$$(A\gamma) = 20$$

We require the limits, if any, of (BC).

Let us note first of all that we are given 6 class-frequencies (including N). If we knew two more, independent of these 6, the problem would be completely determinate, for we should have 2^3 class-frequencies.

Let us therefore put

$$(\alpha\beta\gamma) = x$$
$$(ABC) = y$$

We can then solve for the ultimate class-frequencies and get

$$(AB\gamma) = 45 - y$$
$$(A\beta C) = 30 - y$$
$$(\alpha BC) = x - 15$$
$$(A\beta\gamma) = y - 25$$
$$(\alpha B\gamma) = 30 - x$$
$$(\alpha\beta C) = 35 - x$$

The condition that these must be non-negative gives us conditions on x and y. In fact, from (αBC) and $(\alpha B\gamma)$ we get

$$15 \ngtr x \ngtr 30$$

and from $(A\beta C)$ and $(A\beta\gamma)$,

$$25 \ngtr y \ngtr 30$$

the conditions from the other frequencies being included in these limits to x and y.

Now

$$(BC) = (ABC) + (\alpha BC)$$
$$= y + x - 15 \qquad .$$

and hence, from the limits to x and y,

$$25 \ngtr (BC) \ngtr 45$$

Consequently, the percentage of the population 45 years old or more (50 per cent. of the total population) who are wage-earners lies between 50 and 90 per cent.

It is worth while examining whether these limits are the narrowest possible which can be assigned with the available data ; and it is easy to see that they are. For if $x = 15$ and $y = 25$, $(BC) = 25$; and if $x = 30$ and $y = 30$, $(BC) = 45$. There is nothing in the conditions of the problem to prevent x and y, and hence (BC), from reaching the limiting values, and thus no narrowing of the limits is possible.

SUMMARY.

1. The necessary and sufficient condition for the consistence of a set of independent class-frequencies relating to a particular universe is that no ultimate class-frequency which may be calculated from them is negative.

2. In view of the practical importance of the positive class-frequencies, the form of the consistence conditions is expressed solely in terms of such frequencies.

3. The conditions may be applied to the examination of inaccurate or incomplete data. For the latter they may allow us to assign limits to an unknown class-frequency.

EXERCISES.

2.1. (For this and similar estimates *cf.* "Report by Miss Collet on the Statistics of Employment of Women and Girls" [C,—7564], 1894.)   If, in the urban district of Bury, 817 per thousand of the women between 20 and 25 years of age were returned as "occupied" at the census of 1891, and 263 per thousand as married or widowed, what is the lowest proportion per thousand of the married or widowed that must have been occupied?

2.2. If, in a series of houses actually invaded by smallpox, 70 per cent. of the inhabitants are attacked and 85 per cent. have been vaccinated, what is the lowest percentage of the vaccinated that must have been attacked?

2.3. Given that 50 per cent. of the inmates of a workhouse are men, 60 per cent. are "aged" (over 60), 80 per cent. non-able-bodied, 35 per cent. aged men, 45 per cent. non-able-bodied men, and 42 per cent. non-able-bodied and aged, find the greatest and least possible proportions of non-able-bodied aged men.

2.4. (Material from ref. (69).)   The following are the proportions per 10,000 of boys observed for certain classes of defects amongst a number of school-children. A = development defects, B = nerve signs, D = mental dulness.

$$
\begin{array}{ll}
N = 10,000 & (D) = 789 \\
(A) = 877 & (AB) = 338 \\
(B) = 1,086 & (BD) = 455
\end{array}
$$

Show that some dull boys do not exhibit development defects, and state how many at least do not do so.

2.5. The following are the corresponding figures for girls:—

$$
\begin{array}{ll}
N = 10,000 & (D) = 689 \\
(A) = 682 & (AB) = 248 \\
(B) = 850 & (BD) = 363
\end{array}
$$

Show that some defectively developed girls are not dull, and state how many at least must be so.

2.6. Take the syllogism "All A's are B's, all B's are C's, therefore all A's are C's," express the premises in terms of the notation of the preceding chapters, and deduce the conclusion by the use of the general conditions of consistence.

2.7. Do the same for the syllogism "All A's are B's, no B's are C's, therefore no A's are C's."

2.8. Given that $(A) = (B) = (C) = \frac{1}{2}N$, and that $(AB)/N = (AC)/N = p$, find what must be the greatest and least values of p in order that we may infer that $(BC)/N$ exceeds any given value, say q.

2.9. Show that if

$$
\frac{(A)}{N} = x \qquad \frac{(B)}{N} = 2x \qquad \frac{(C)}{N} = 3x
$$

and

$$
\frac{(AB)}{N} = \frac{(AC)}{N} = \frac{(BC)}{N} = y
$$

the value of neither x nor y can exceed $\frac{1}{4}$.

2.10. A market investigator returns the following data. Of 1000 people consulted, 811 liked chocolates, 752 liked toffee and 418 liked boiled sweets; 570 liked chocolates and toffee, 356 liked chocolates and boiled sweets and 348 liked toffee and boiled sweets; 297 liked all three.   Show that this information as it stands must be incorrect.

2.11. (Imaginary data.) 50 per cent. of the imports of barley into a country come from the Dominions; 80 per cent. of the total imports go to brewing;

75 per cent. of the imports are grown in the Northern hemisphere; 80 per cent. of Northern-grown barley goes to brewing; 100 per cent. of foreign Southern-grown barley goes to stock-feeding. Show that the foreign Northern-grown barley which goes to brewing cannot be less than 30 per cent. nor more than 60 per cent. of the total imports.

(It is assumed that brewing and stock-feeding are the only two uses to which imported barley is put.)

2.12. A penny is tossed three times and the results, heads and tails, noted. The process is continued until there are 100 sets of threes. In 69 cases heads fell first, in 49 cases heads fell second, and in 53 cases heads fell third. In 33 cases heads fell both first and second, and in 21 cases heads fell both second and third. Show that there must have been at least 5 occasions on which heads fell three times, and that there could not have been more than 15 occasions on which tails fell three times, though there need not have been any.

CHAPTER 3.

ASSOCIATION OF ATTRIBUTES.

Independence.

3.1. If there is no sort of relationship of any kind between two attributes A and B, we expect to find the same proportion of A's amongst the B's as amongst the not-B's. We may anticipate, for instance, the same proportion of abnormally wet seasons in leap years as in ordinary years, the same proportion of male to total births when the moon is waxing as when it is waning, the same proportion of heads whether a coin be tossed with the right hand or the left.

Two such unrelated attributes may be termed **independent**, and we have accordingly as the **criterion of independence** for A and B:

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \qquad . \qquad . \qquad . \qquad . \quad (3.1)$$

If this relation hold good, the corresponding relations

$$\frac{(aB)}{(B)} = \frac{(a\beta)}{(\beta)}$$

$$\frac{(AB)}{(A)} = \frac{(aB)}{(a)}$$

$$\frac{(A\beta)}{(A)} = \frac{(a\beta)}{(a)}$$

must also hold. For it follows at once from (3.1) that

$$\frac{(B)-(AB)}{(B)} = \frac{(\beta)-(A\beta)}{(\beta)}$$

that is,

$$\frac{(aB)}{(B)} = \frac{(a\beta)}{(\beta)}$$

and the other two identities may be similarly deduced.

The student may find it easier to grasp the nature of the relations stated if the frequencies are supposed grouped into a table with two rows and two columns, thus :

Attribute.	Attribute.		Total.
	B	β	
A	(AB)	$(A\beta)$	(A)
a	(aB)	$(a\beta)$	(a)
Total	(B)	(β)	N

34

Equation (3.1) states a certain equality for the columns ; if this holds good, the corresponding equation

$$\frac{(AB)}{(A)} = \frac{(aB)}{(a)}$$

must hold for the rows, and so on.

Forms of the Criterion of Independence.

3.2. The criterion may, however, be put into a somewhat different and theoretically more convenient form. The equation (3.1) expresses (AB) in terms of (B), (β) and a second-order frequency $(A\beta)$; eliminating this second-order frequency we have :

$$\frac{(AB)}{(B)} = \frac{(AB)+(A\beta)}{(B)+(\beta)} = \frac{(A)}{N}$$

i.e. in words, " the proportion of A's amongst the B's is the same as in the universe at large." The student should learn to recognise this equation at sight in any of the forms :

$$
\left.
\begin{aligned}
\frac{(AB)}{(B)} &= \frac{(A)}{N} \qquad &(a)\\[6pt]
\frac{(AB)}{(A)} &= \frac{(B)}{N} \qquad &(b)\\[6pt]
(AB) &= \frac{(A)(B)}{N} \qquad &(c)\\[6pt]
\frac{(AB)}{N} &= \frac{(A)}{N} \cdot \frac{(B)}{N} \qquad &(d)
\end{aligned}
\right\}
\qquad . \qquad . \qquad . \quad (3.2)
$$

The equation (d) gives the important fundamental rule : *If the attributes A and B are independent, the proportion of AB's in the universe is equal to the proportion of A's multiplied by the proportion of B's.*

The advantage of the forms (3.2) over the form (3.1) is that they give expressions for the second-order frequency in terms of the frequencies of the first order and the whole number of observations alone ; the form (3.1) does not.

Example 3.1.—If there are 144 A's and 384 B's in 1024 observations, how many AB's will there be, A and B being independent ?

$$\frac{144 \times 384}{1024} = 54$$

There will therefore be 54 AB's.

Example 3.2.—If the A's are 60 per cent., the B's 35 per cent., of the whole number of observations, what must be the percentage of AB's in order that we may conclude that A and B are independent ?

$$\frac{60 \times 35}{100} = 21$$

and therefore there must be 21 per cent. (more or less closely, *cf.* **3.8** and **3.9** below) of AB's in the universe to justify the conclusion that A and B are independent.

3.3. It follows from **3.1** that if the relation (3.2) holds for any one of the four second-order frequencies, *e.g.* (AB), similar relations must hold for the remaining three. Thus we have directly from (3.1):

$$\frac{(A\beta)}{(\beta)} = \frac{(AB)+(A\beta)}{(B)+(\beta)} = \frac{(A)}{N}$$

giving

$$(A\beta) = \frac{(A)(\beta)}{N}$$

and so on. This is seen at once to be true on consideration of the fourfold table on page 34. For if (AB) takes the value $(A)(B)/N$, $(A\beta)$ must take the value $(A)(\beta)/N$ to keep the total of the row equal to (A), and so on for the other rows and columns. The fourfold table in the case of independence must in fact have the form:

Attribute.	Attribute.		Total.
	B	β	
A	$(A)(B)/N$	$(A)(\beta)/N$	(A)
a	$(a)(B)/N$	$(a)(\beta)/N$	(a)
Total	(B)	(β)	N

Example 3.3.—In Example 3.1 above, what would be the number of $a\beta$'s, A and B being independent?

$$(a) = 1024 - 144 = 880$$
$$(\beta) = 1024 - 384 = 640$$
$$\therefore \quad (a\beta) = \frac{880 \times 640}{1024} = 550$$

3.4. Finally, the criterion of independence may be expressed in yet a third form, viz. in terms of the second-order frequencies alone. If A and B are independent, it follows at once from the preceding section that

$$(AB)(a\beta) = \frac{(A)(B)(a)(\beta)}{N^2}$$

And evidently $(aB)(A\beta)$ is equal to the same fraction.

Therefore

$$\left. \begin{array}{ll} (AB)(a\beta) = (aB)(A\beta) & (a) \\[6pt] \dfrac{(AB)}{(aB)} = \dfrac{(A\beta)}{(a\beta)} & (b) \\[6pt] \dfrac{(AB)}{(A\beta)} = \dfrac{(aB)}{(a\beta)} & (c) \end{array} \right\} \qquad . \qquad . \qquad . \quad (3.8)$$

The equation (b) may be read: " The ratio of A's to a's amongst the B's is equal to the ratio of A's to a's amongst the β's," and (c) similarly.

This form of criterion is a convenient one if all the four second-order frequencies are given, enabling one to recognise almost at a glance whether or not the two attributes are independent.

Example 3.4.—If the second-order frequencies have the following values, are A and B independent or not?

$$(AB) = 110 \qquad (aB) = 90 \qquad (A\beta) = 290 \qquad (a\beta) = 510$$

Clearly

$$(AB)(a\beta) > (aB)(A\beta)$$

so A and B are not independent.

Association.

3.5. Suppose now that A and B are not independent, but related in some way or other, however complicated.

Then if

$$(AB) > \frac{(A)(B)}{N}$$

A and B are said to be **positively associated**, or sometimes simply **associated**. If, on the other hand,

$$(AB) < \frac{(A)(B)}{N} \ .$$

A and B are said to be **negatively associated** or, more briefly, **disassociated**.

The student should carefully note that in statistics the word " association " has a technical meaning different from the one current in ordinary speech. In common language one speaks of A and B as being " associated " if they appear together in a number of cases. But in statistics A and B are associated only if they appear together in a greater number of cases than is to be expected if they are independent. Thus, if we consider means of land transport as dichotomised into road and rail travel, we may say, in the customary use of the term, that road transport is associated with speed. But it does not follow that the two are statistically associated, because rail transport may equally be associated with speed and, in fact, the attribute speed may be independent of the means of travel in these two manners.

Association, therefore, cannot be inferred from the mere fact that *some* A's are B's, however great the proportion : this principle is fundamental and should always be borne in mind.

Complete Association and Disassociation.

3.6. We have now to consider in what circumstances we may regard the association of two attributes as **complete**. Two courses are open to us. Either we may say that for complete association all A's must be B's *and* all B's must be A's, in which case it must follow that the A's and the B's occur in the universe in equal numbers ; or we may adopt a rather wider meaning and say that all A's are B's *or* all B's are A's,

according to whether the A's or the B's are in the minority. Similarly, complete disassociation may be taken either as the case when no A's are B's *and* no a's are β's, or more widely as the case when either of these statements is true.

We shall adopt the wider definition in the sequel. Thus two attributes are **completely associated** if one of them cannot occur without the other, though the other may occur without the one.

Measurement of Intensity of Association.

3.7. It follows from the foregoing that if two attributes are completely associated, (AB) must be equal to (A) or (B), whichever is the smaller. If they are completely disassociated, (AB) must be equal to zero or to $(A) + (B) - N$, whichever is the greater. (AB) must in general lie between these two limits. We may thus regard the divergence of (AB) from the " independence " value $(A)(B)/N$ towards the limiting value in either direction as indicating the *intensity* of association or disassociation, so that we may speak of attributes as being *more* or *less, highly* or *slightly,* associated. This conception of degrees of association quantitatively expressible is important, and we return in a later section to consider the formulæ which may be used to measure such degrees.

Sampling Fluctuations.

3.8. When the association is very slight, *i.e.* where (AB) only differs from $(A)(B)/N$ by a few units or by a small proportion, it may be that such association is not really significant of any definite relationship. To give an illustration, suppose that a coin is tossed a number of times, and the tosses noted in pairs ; then 100 pairs may give such results as the following (taken from an actual record) :—

First toss heads and second heads . . .	26
,, ,, ,, tails . . .	18
First toss tails and second heads . . .	27
,, ,, ,, tails . . .	29

If we use A to denote " heads " in the first toss, B " heads " in the second, we have from the above $(A) = 44$, $(B) = 53$. Hence $(A)(B)/N = \dfrac{44 \times 53}{100} = 23 \cdot 32$, while actually (AB) is 26. Hence there is a positive association, in the given record, between the result of the first throw and the result of the second. But it is fairly certain, from the nature of the case, that such association cannot indicate any real connection between the results of the two throws ; it must therefore be due merely to such a complex system of causes, impossible to analyse, as leads, for example, to differences between small samples drawn from the same material. The conclusion is confirmed by the fact that, of a number of such records, some give a positive association (like the above), but others a negative association.

3.9. An event due, like the above occurrence of positive association, to an extremely complex system of causes of the general nature of which we are aware, but of the detailed operation of which we are ignorant, is sometimes said to be due to *chance*, or better to the **chances** or **fluctuations of sampling.**

A little consideration will suggest that such associations due to the fluctuations of sampling must be met with in all classes of statistics. To quote, for instance, from **3.1**, two illustrations there given of independent attributes, we know that in any *actual* record we would not be likely to find *exactly* the same proportion of abnormally wet seasons in leap years as in ordinary years, nor *exactly* the same proportion of male births when the moon is waxing as when it is waning. But so long as the divergence from independence is not well marked we must regard such attributes as practically independent, or dependence as at least unproved.

The discussion of the question, how great the divergence must be before we can consider it as " well marked," must be postponed to the chapters dealing with the theory of sampling. At present the attention of the student can only be directed to the existence of the difficulty, and to the serious risk of interpreting a " chance association " as physically significant.

The Choice of a Suitable Form for Testing Association.

3.10. The definition of **3.5** suggests that we are to test the existence or the intensity of association between two attributes by a comparison of the actual value of (AB) with its independence value (as it may be termed) $(A)(B)/N$. The procedure is from the theoretical standpoint perhaps the most natural, but it is more usual, and is simplest and best in practice, to compare *proportions*, *e.g.* the proportion of A's amongst the B's with the proportion amongst the β's. Such proportions are usually expressed in the form of percentages or proportions per thousand.

It will be evident from **3.1** and **3.2** that a large number of such comparisons are available for the purpose, and the question arises, therefore, which is the best comparison to adopt ?

3.11. Two principles should decide this point : (1) of any two comparisons, that is the better which brings out the more clearly the degree of association ; (2) of any two comparisons, that is the better which illustrates the more important aspect of the problem under discussion.

The first condition at once suggests that comparisons of the form

$$\frac{(AB)}{(B)} > \frac{(A\beta)}{(\beta)} \qquad . \qquad . \qquad . \qquad . \quad (3.4)$$

are better than comparisons of the form

$$\frac{(AB)}{(B)} > \frac{(A)}{N} \qquad . \qquad . \qquad . \qquad . \quad (3.5)$$

For it is evident that if most of the objects or individuals in the universe are B's, *i.e.* if $(B)/N$ approaches unity, $(AB)/(B)$ will necessarily approach $(A)/N$ even though the difference between $(AB)/(B)$ and $(A\beta)/(\beta)$ is considerable. The second form of comparison may therefore be misleading.

Setting aside, then, comparisons of the general form (3.5), the question remains whether to apply the comparison of the form (3.4) to the rows or the columns of the table, if the data are tabulated as on page 34. This question must be decided with reference to the second principle, *i.e.* with regard to the more important aspect of the problem under discussion,

the exact question to be answered, or the hypothesis to be tested, as illustrated by the examples below. Where no *definite* question has to be answered or hypothesis tested both pairs of proportions may be tabulated, as in Example 3.6.

Example 3.5.—Association between inoculation against cholera and exemption from attack. (Data from Greenwood and Yule, Table III, ref. (74).)

	Not attacked.	Attacked.	Total.
Inoculated . .	276	3	279
Not inoculated . .	473	66	539
Total . .	749	69	818

Here the important question is, How far does inoculation protect from attack ? The most natural comparison is therefore—

Percentage of inoculated who were not attacked . . 98·9
„   not inoculated   „   „   . . 87·8

Or we might tabulate the complementary proportions—

Percentage of inoculated who were attacked . . 1·1
„   not inoculated   „   „   . . 12·2

Either comparison brings out simply and clearly the fact that *inoculation* and *exemption from attack* are *positively associated* (*inoculation* and *attack* negatively associated).

We are making above a comparison by rows in the notation of the table on page 34, comparing $(AB)/(A)$ with $(aB)/(a)$, or $(A\beta)/(A)$ with $(a\beta)/(a)$. A comparison by columns, *e.g.* $(AB)/(B)$ with $(A\beta)/(\beta)$, would serve equally to indicate whether there was any appreciable association, but would not answer directly the particular question we have in mind :

Percentage of not-attacked who were inoculated . . 36·8
„   attacked   „   „   . . 4·3

Example 3.6.—Deaf-mutism and Imbecility. (Material from Census of 1901. Summary Tables. (Cd. 1523).)

Total population of England and Wales . . 32,528,000
Number of the imbecile (or feeble-minded) . 48,882
Number of deaf-mutes 15,246
Number of imbecile deaf-mutes . . . 451

Required, to find whether deaf-mutism is associated with imbecility.

We may denote the number of the imbecile by (A), of deaf-mutes by (B). A comparison of $(AB)/(B)$ with $(A)/N$ or of $(AB)/(A)$ with $(B)/N$ may very well be used in this case, seeing that $(A)/N$ and $(B)/N$ are both small. The question whether to give the preference to the first or the second comparison depends on the nature of the investigation we wish to

make. If it is desired to exhibit the conditions among deaf-mutes the first may be used :

Proportion of imbeciles among deaf-mutes $= (AB)/(B)$ . . } 29·6 per thousand

Proportion of imbeciles in the whole population $= (A)/N$ . . . } 1·5 ,,

If, on the other hand, it is desired to exhibit the conditions amongst the imbecile, the second will be preferable:

Proportion of deaf-mutes amongst the imbecile $= (AB)/(A)$ . . } 9·2 per thousand

Proportion of deaf-mutes in the whole population $= (B)/N$. } 0·5 ,,

Either comparison exhibits very clearly that there exists an association between the attributes. It may be pointed out, however, that census data as to such infirmities are very untrustworthy.

Example 3.7.—Eye-colour of father and son (material due to Sir Francis Galton, as given by Professor Karl Pearson, *Phil. Trans.*, A, vol. 195, 1900, p. 138; the classes 1, 2 and 3 of the memoir treated as "light ").

Fathers with light eyes and sons with light eyes (AB) . . 471

,, ,, ,, ,, not light ,, $(A\beta)$ . . 151

,, not light ,, ,, light ,, (aB) . . 148

,, ,, ,, ,, not light ,, $(a\beta)$ . . 230

Required to find whether the colour of the son's eyes is associated with that of the father's. In cases of this kind the father is reckoned once for each son ; *e.g.* a family in which the father was light-eyed, two sons light-eyed and one not, would be reckoned as giving two to the class AB and one to the class $A\beta$.

The best comparison here is—

Percentage of light-eyed amongst the sons of light-eyed fathers . . . . } 76 per cent.

Percentage of light-eyed amongst the sons of not-light-eyed fathers . . . } 39 ,,

But the following is equally valid :—

Percentage of light-eyed amongst the fathers of light-eyed sons . . . } 76 per cent.

Percentage of light-eyed amongst the fathers of not-light-eyed sons . . } 40 ,,

The reason why the former comparison is preferable is that we usually wish to estimate the character of offspring from that of the parents, and not *vice versa*. Both modes of statement, however, indicate equally clearly that there is considerable resemblance between father and son.

Example 3.8.—Association between inoculation against cholera and exemption from attack, five separate epidemics (*cf.* Example 3.5, data from Tables IX, X, XXVIII, XXIX, XXXI of ref. (74)).

	Not attacked.	Attacked.	Total.
Inoculated	192	4	196
Not inoculated . . .	113	34	147
Total . . .	305	38	343

	Not attacked.	Attacked.	Total.
Inoculated	5,751	27	5,778
Not inoculated . . .	6,351	198	6,549
Total . . .	12,102	225	12,327

	Not attacked.	Attacked.	Total.
Inoculated	4,087	5	4,092
Not inoculated . . .	113,856	1,144	115,000
Total . . .	117,943	1,149	119,092

	Not attacked.	Attacked.	Total.
Inoculated	8,332	8	8,340
Not inoculated . . .	84,444	556	85,000
Total . . .	92,776	564	93,340

	Not attacked.	Attacked.	Total.
Inoculated	4,870	5	4,875
Not inoculated . . .	153,096	904	154,000
Total . . .	157,966	909	158,875

With the table of Example 3.5 the above give data for six separate epidemics, in all of which the same method of inoculation appears to have been used : the data refer to natives only, and the numbers of observations are sufficiently large to reduce " fluctuations of sampling " within reasonably narrow limits. The proportions not attacked are as follows : —

Proportion not Attacked.			
	Not Inoculated.	Inoculated.	Difference.
1 . . .	0·8776	0·9892	0·1116
2 . . .	0·7687	0·9796	0·2109
3 . . .	0·9698	0·9953	0·0255
4 . . .	0·9901	0·9988	0·0087
5 . . .	0·9935	0·9990	0·0055
6 . . .	0·9941	0·9990	0·0049

In each case *inoculation* and *exemption from attack* are positively associated, but it will be seen that the several proportions, and the differences between them, vary considerably. Evidently in a very mild

epidemic this difference can only be small, and the question arises how far the data for the separate epidemics can be said to be consistent in their indication of the " efficiency " of the inoculation. This is not a simple question to answer : the more advanced student is referred to the discussion in the original.

The Symbols $(AB)_0$ and δ.

3.12. The values that the four second-order frequencies take in the case of independence, viz.

$$\frac{(A)(B)}{N}, \quad \frac{(a)(B)}{N}, \quad \frac{(A)(\beta)}{N}, \quad \frac{(a)(\beta)}{N}$$

are of such great theoretical importance, and of so much use as reference-values for comparing with the actual values of the frequencies (AB), (aB), $(A\beta)$ and $(a\beta)$, that it is often desirable to employ single symbols to denote them. We shall use the symbols

$$(AB)_0 = \frac{(A)(B)}{N} \qquad (a\beta)_0 = \frac{(a)(\beta)}{N}$$

$$(aB)_0 = \frac{(a)(B)}{N} \qquad (A\beta)_0 = \frac{(A)(\beta)}{N}$$

If δ denote the excess of (AB) over $(AB)_0$, then, in order to keep the totals of rows and columns constant, the general table (*cf.* the table for the case of independence on page 36) must be of the form

Attribute.	Attribute.		Total.
	B	β	
A	$(AB)_0 + \delta$	$(A\beta)_0 - \delta$	(A)
a	$(aB)_0 + \delta$	$(a\beta)_0 + \delta$	(a)
Total	(B)	(β)	N

Therefore, quite generally we have :

$$(AB) - (AB)_0 = (a\beta) - (a\beta)_0 = (A\beta)_0 - (A\beta) = (aB)_0 - (aB) = \delta$$

3.13. The value of this common difference δ may be expressed in a form that is useful to note. We have by definition :

$$\delta = (AB) - (AB)_0 = (AB) - \frac{(A)(B)}{N}$$

Bring the terms on the right to a common denominator, and express all the frequencies of the numerator in terms of those of the second order ; then we have :

$$\delta = \frac{1}{N} \left\{ \begin{array}{l} (AB)[(AB) + (aB) + (A\beta) + (a\beta)] \\ - [(AB) + (A\beta)][(AB) + (aB)] \end{array} \right\}$$

$$= \frac{1}{N} \{ (AB)(a\beta) - (aB)(A\beta) \}$$

That is to say, the common difference is equal to $1/N$th of the difference of the " cross-products " $(AB)(a\beta)$ and $(a\tilde{B})(A\beta)$.

It is evident that the difference of the cross-products may be very large if N be large, although δ is really very small.   In using the difference of the cross-products to test mentally the sign of the association in a case where all the four second-order frequencies are given, this should be remembered : the difference should be compared with N, or it will be liable to suggest a higher degree of association than actually exists.

Example 3.9.—The following data were observed for hybrids of *Datura* (W. Bateson and Miss Saunders, Report to the Evolution Committee of the Royal Society, 1902) :—

Flowers violet, fruits prickly (AB)	.	.	47
,,	,,	smooth $(A\beta)$	.	.	12
Flowers white,	,,	prickly (aB)	.	.	21
,,	,,	smooth $(a\beta)$	.	.	3

Investigate the association between colour of flower and character of fruit.

Since $3 \times 47 = 141$, $12 \times 21 = 252$, *i.e.* $(AB)(a\beta) < (aB)(A\beta)$, there is clearly a negative association ; $252 - 141 = 111$, and at first sight this considerable difference is apt to suggest a considerable disassociation.   But $\delta = 111/83 = 1\cdot3$ only, and forms a small proportion of the frequency, so that in point of fact the disassociation is small, so small that no stress can be laid on it as indicating anything but a fluctuation of sampling.   Working out the percentages we have :

Percentage of violet-flowered plants with prickly fruits . . . . . $\Big\}$ 80 per cent.

Percentage of white-flowered plants with prickly fruits . . . . . $\Big\}$ 87   ,,

Coefficient of Association.

3.14.   In the previous examples we have judged the association by comparing the class-frequencies with those which would exist if the data were given by independent attributes, and we can form a rough idea of the strength of the association by examining the extent of the difference. This is sufficient for almost all practical purposes, although, if the data are likely to be affected seriously by fluctuations of random sampling, some test of the significance of the difference is also necessary.   Apart from this question, however, it is sometimes convenient to measure the intensities of the associations by means of a coefficient.

It is clearly convenient if such a coefficient can be devised as to be zero if the attributes are independent, $+1$ if they are completely associated and -1 if they are completely disassociated.

3.15.   Many such coefficients may be devised, but perhaps the simplest possible (though not necessarily the most advantageous) is the expression—

$$Q = \frac{(AB)(a\beta) - (A\beta)(aB)}{(AB)(a\beta) + (A\beta)(aB)}$$

$$= \frac{N\delta}{(A\bar{B})(a\beta) + (A\beta)(aB)}$$

where δ is the symbol used in **3.12** and **3.13** for the difference $(AB) - (AB)_0$. It is evident that Q is zero when the attributes are independent, for then δ is zero : it takes the value $+1$ when there is complete association, for then the second term in both numerator and denominator of the first form of the expression is zero : similarly it is -1 where there is complete disassociation, for then the first term in both numerator and denominator is zero. Q may accordingly be termed a **coefficient of association.** As illustrations of the values it will take in certain cases, the association between deaf-mutism and imbecility, on the basis of the English census figures (Example 3.6), is $+0.91$; between light eye-colour in father and in son (Example 3.7), $+0.66$; between colour of flower and prickliness of fruit in *Datura* (Example 3.9), -0.28—a disassociation which, however, as already stated, is probably of no practical significance and due to mere fluctuations of sampling.

The student should note that if all the terms containing A are multiplied by a constant, the value of Q is unaltered. Similarly for a, B and β. Hence Q is independent of the relative proportions of A's and a's in the data. This property is important, and renders such a measure of association specially adapted to cases in which the proportions are arbitrary (*e.g.* experiments). A form possessing the same property but certain marked advantages over Q is suggested in ref. (80).

3.16. The coefficient is only mentioned here to direct the attention of the student to the possibility of forming such a measure of association, a measure which serves a similar purpose in the case of attributes to that served by certain other coefficients in the cases of manifold classification (*cf.* Chap. 5) and of variables (*cf.* Chap. 11, and the references to Chaps. 11, 12 and 13). For further illustrations of the use of this coefficient the reader is referred to ref. (78); for a modified form of the coefficient, possessing the same properties but certain advantages, to ref. (80); and for a mode of deducing another coefficient, based on theorems in the theory of variables, which has come into more general use, though in the opinion of the present writers its use is of doubtful advantage, to ref. (76). Reference should also be made to the coefficient described in **13.25.** The question of the best coefficient to use as a measure of association is one on which statisticians differ: for a discussion the student is referred to refs. (74), (77) and (80).

A Necessary Caution.

3.17. In concluding this chapter, it may be well to repeat, for the sake of emphasis, that the mere fact of 80, 90 or 99 per cent. of A's being B's implies nothing as to the association of A with B; in the absence of information, we can but assume that 80, 90 or 99 per cent. of a's may also be B's. In order to apply the criterion of independence for two attributes A and B, it is necessary to have information concerning a's and β's as well as A's and B's, or concerning a universe that includes both a's and A's, β's and B's. Hence an investigation as to the causal relations of an attribute A must not be confined to A's, but must be extended to a's (unless, of course, the necessary information as to a's is already obtainable) : no *comparison* is otherwise possible. It would be no use to obtain with great pains the result (*cf.* Example 3.6) that 29.6 per thousand of deaf-mutes were imbecile unless we knew that the proportion of imbeciles in the

whole population was only 1·5 per thousand; nor would it contribute anything to our knowledge of the heredity of deaf-mutism to find out the proportion of deaf-mutes amongst the offspring of deaf-mutes unless the proportions amongst the offspring of normal individuals were also investigated or known.

SUMMARY.

1. Two attributes are independent if the proportion of A's among the B's is the same as the proportion among the not-B's.

2. This definition can be expressed symbolically in numerous forms, in terms of either first-order or second-order frequencies. The form in which the data are given, and the question which is to be answered, determine which form is to be employed in any particular case.

3. Attributes which are not independent are said to be positively associated if

$$(AB) > \frac{(A)(B)}{N}$$

and negatively associated if

$$(AB) < \frac{(A)(B)}{N}.$$

4. The statistical meaning of the word " association " is different from the meaning ascribed to it in ordinary language.

5. Before association may be said to indicate a definite relation between the attributes, it is necessary to be satisfied that the divergence from independence is not due to fluctuations of sampling.

6. The divergence of the actual frequency from the " independence " frequency is denoted by the symbol δ, and hence

$$\delta = (AB) - \frac{(A)(B)}{N}$$

7. The coefficient of association is defined by

$$Q = \frac{N\delta}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

It is zero if the attributes are independent, $+1$ if they are completely associated and -1 if they are completely disassociated. There are, however, other forms of coefficient more advantageous in certain cases (ref. (80)).

EXERCISES.

3.1. At the census of England and Wales in 1901 there were (to the nearest 1000) 15,729,000 males and 16,799,000 females; 3497 males were returned as deaf-mutes from childhood, and 3072 females.

State proportions exhibiting the association between deaf-mutism from childhood and sex. How many of each sex for the same total number would have been deaf-mutes if there had been no association?

3.2. Show, as briefly as possible, whether A and B are independent, positively associated or negatively associated in each of the following cases:—

(a)	$N = 5000$	$(A) = 2350$	$(B) = 3100$	$(AB) = 1600$
(b)	$(A) = 490$	$(AB) = 294$	$(a) = 570$	$(aB) = 380$
(c)	$(AB) = 256$	$(aB) = 768$	$(A\beta) = 48$	$(a\beta) = 144$

3.3. (Figures derived from Darwin's "*Cross- and Self-fertilisation of Plants.*") The table below gives the numbers of plants of certain species that were above or below the average height, stating separately those that were derived from cross-fertilised and from self-fertilised parentage. Investigate the association between height and cross-fertilisation of parentage, and draw attention to any special points you notice.

Species.	Parentage Cross-fertilised. Height—		Parentage Self-fertilised. Height—	
	Above Average.	Below Average.	Above Average.	Below Average.
Ipomœa purpurea	63	10	18	55
Petunia violacea	61	16	13	64
Reseda lutea	25	7	11	21
Reseda odorata	39	16	25	30
Lobelia fulgens	17	17	12	22

3.4. (Figures from same source as Example 3.7, p. 41, but material differently grouped; classes 7 and 8 of the memoir treated as "dark.") Investigate the association between darkness of eye-colour in father and son from the following data :—

Fathers with dark eyes and sons with dark eyes			(AB)	.	50
,, ,, ,,	not-dark eyes		$(A\beta)$	.	79
Fathers with not-dark eyes and sons with dark eyes			(aB)	.	89
,, ,, ,,	not-dark eyes	$(a\beta)$		.	782

Also tabulate for comparison the frequencies that would have been observed had there been no heredity, *i.e.* the values of $(AB)_0$, $(A\beta)_0$, etc.

3.5. (Figures from same source as above.) Investigate the association between eye-colour of husband and eye-colour of wife ("assortative mating") from the data given below.

Husbands with light eyes and wives with light eyes			(AB)	.	309
,, ,, ,,	not-light eyes		$(A\beta)$	.	214
Husbands with not-light eyes and wives with light eyes			(aB)	.	132
,, ,, ,,	not-light eyes	$(a\beta)$		.	119

Also tabulate for comparison the frequencies that would have been observed had there been strict independence between eye-colour of husband and eye-colour of wife, *i.e.* the values of $(AB)_0$, etc., as in Exercise 3.4.

3.6. (Figures from the *Census of England and Wales, 1891*, vol. 3: the data cannot be regarded as trustworthy.) The figures given below show the number of males in successive age-groups, together with the number of the blind (A), of the mentally deranged (B) and the blind mentally deranged (AB). Trace the association between blindness and mental derangement from childhood to old age, tabulating the proportions of insane amongst the whole population and amongst the blind, and also the association coefficient Q of 3.15. Give a short verbal statement of your results.

	5–	15–	25–	35–	45–	55–	65–	75 and upwards.
N	3,304,230	2,712,521	2,089,010	1,611,077	1,191,789	770,124	444,896	161,692
(A)	844	1,184	1,165	1,501	1,752	1,905	1,932	1,701
(B)	2,820	6,226	8,482	9,214	8,187	5,799	3,412	1,098
(AB)	17	19	19	31	32	34	22	9

3.7. Show that if

$$
\begin{array}{cccc}
(AB)_1 & (aB)_1 & (A\beta)_1 & (a\beta)_1 \\
(AB)_2 & (aB)_2 & (A\beta)_2 & (a\beta)_2
\end{array}
$$

be two aggregates corresponding to the same values of (A), (B), (a) and (β),

$$
(AB)_1 - (AB)_2 = (aB)_2 - (aB)_1 = (A\beta)_2 - (A\beta)_1 = (a\beta)_1 - (a\beta)_2
$$

3.8. Show that if

$$
\delta = (AB) - (AB)_0
$$

$$
(AB)^2 + (a\beta)^2 - (aB)^2 - (A\beta)^2 = [(A) - (a)][(B) - (\beta)] + 2N\delta
$$

3.9. The existence of association may be tested either by comparison of proportions (*e.g.* $(AB)/(B)$ with $(A\beta)/(\beta)$), as in 3.10 and 3.11, or by the value of δ, as in 3.12 and 3.13. Show that

$$
\delta = \frac{(B)(\beta)}{N}\left\{\frac{(AB)}{(B)} - \frac{(A\beta)}{(\beta)}\right\}
$$

$$
= \frac{(A)(a)}{(N)}\left\{\frac{(AB)}{(A)} - \frac{(aB)}{(a)}\right\}
$$

3.10. Spence and Charles, in *An Investigation into the Health and Nutrition of Certain of the Children of Newcastle-on-Tyne between the Ages of One and Five Years* (City and Council of Newcastle-on-Tyne, February 1934), compared two groups of children, one belonging to the professional classes, 125 in number, and the other belonging to the labouring classes, 124 in number. They found the following results:—

	Poor Children. Per cent.	Well-to-do Children. Per cent.
Below normal weight . . .	55	13
Above normal weight . . .	11	48

Find the coefficient of association between the weight of the children and their social status.

3.11. (Data from the *Report on the Spahlinger Experiments in Northern Ireland*, 1931–1934, H.M. Stationery Office, 1935.) In experiments on the immunisation of cattle from tuberculosis the following results were secured:—

	Cattle.		
	Died of Tuberculosis or very seriously affected.	Unaffected or only slightly affected.	Total.
Inoculated with vaccine . .	6	13	19
Not inoculated or inoculated with control media.	8	3	11
Total . .	14	16	30

(The cattle were first inoculated with protective vaccine and then deliberately infected with serious quantities of tubercle germs.)

Find the coefficient of association between inoculation and exemption from serious tuberculosis.

3.12. Criticise the following argument: "Nearly all the A's are B's, and therefore A and B must be associated," and state what suppressed premises would justify it in the following cases:—

"99 per cent. of the people who drink beer die before reaching 100 years of age. Therefore drinking beer is bad for longevity." .

"99 per cent. of the members who voted for the Army Estimates were military officers. Therefore it was unfair to suppose that the voting was unbiassed."

"In every country where the sale of contraceptives is tolerated by the Government the birth-rate is declining. Therefore contraception must exert an influence on the birth-rate."

3.13. Write down in the form of the table of 3.1 the frequency groups when (1) all A's are B's; (2) all B's are A's; (3) all A's are B's and all B's are A's; and the three similar tables when A and B are completely disassociated.

CHAPTER 4.

PARTIAL ASSOCIATION.

Association in Sub-universes.

4.1. In the last chapter we considered the association of two attributes in a universe without regard to whether any information existed about other attributes in the universe. If, however, such information does exist and, say, we can find the frequency-classes of attributes C, D, etc., the question arises, What are the associations of A and B in the sub-universes C, γ, CD, etc. ?

Thus, if A = standard of health and B = consumption of food, the discussion of the previous chapter would enable us to examine whether health and food consumption were associated in any particular universe, say the population of Great Britain. But we might want to go further than this and examine the association between A and B among males, or among the poorer classes, and compare it with the association among females or among the well-to-do classes, respectively. Defining C = males and D = poor, this amounts to examining the associations of A and B in the universes C, γ, D and δ.

4.2. Associations of this kind are of the utmost importance in statistical practice. As instances of the ways in which they arise let us consider the following two illustrations :—

(1) Suppose that we have established, in the manner of the previous chapter, a positive association between inoculation and exemption from smallpox in a universe of persons. It is natural to infer that this association is due to some causal relation between the two attributes and may be expected to recur in the future ; in short, that smallpox is prevented by vaccination.

This rather hasty conclusion might, however, meet an opponent who argues in this way : vaccination is accepted among the well-to-do classes, but is looked on with suspicion by the lower classes. For this and other reasons most of the unvaccinated persons are drawn from the lower classes. But these are precisely the people whom, from the unhygienic conditions under which they live, one would expect to be exposed to infection and who, moreover, being malnourished, would be more likely to contract disease when they were infected. Hence the comparative exemption of the vaccinated persons is not due to the fact that they have been vaccinated, but to the fact that they belong to the well-to-do classes. It is, as it were, an accident that these people also happen to be from a class which favours vaccination.

Denoting *vaccination* by A, *exemption from attack* by B and *hygienic conditions* by C, this argument amounts to saying that the observed

association between A and B is not of itself causally direct, but is due to the associations of both A and B with C.

Now it is clear that this objection could not be lodged if the hygienic conditions among all the members of the universe were the same. If, therefore, we examine the association of A and B in the sub-universe C and still find an association, the supposed argument would be refuted. We are thus led to a consideration of the association in that sub-universe.

(2) As a second example, suppose that an association is noted between the presence of an attribute in the father and the presence in the son, and also between the presence in the grandfather and the presence in the grandson. The question which arises here is: Does the resemblance between grandfather and grandson arise from a kind of hereditary transmission which may, in the common phrase, "skip a generation," or is it merely due to the fact that the grandfather is like the father and the father is like the son?

Denoting the presence of the attribute in the son, father and grandfather by A, B and C, the question is : Is the association between A and C due to associations between A and B, and B and C?

If the association between A and C is observed among all the cases in which the father possesses the attribute or all those in which he does not, and is still sensible, clearly the association between A and C cannot be due to associations between A and B, B and C; hence, as before, to resolve the question we are led to consider the association between A and C in the sub-universes B and β.

4.3. Generally, ambiguity of the type to which we have just referred arises from the fact that the universe of discussion contains not merely objects possessing the third attribute alone, but a mixture of objects with and without it. To meet the requirements of the discussion we have to consider the associations in sub-universes wherein this attribute is entirely absent or entirely present. By this means we can go deeper into the nature of the underlying causes and eliminate certain possible explanations of the type : an association between A and B does not mean that the two are directly related, but only that each is associated with a third attribute C.

Partial Associations.

4.4. The associations between A and B in sub-universes are called **partial associations**, to distinguish them from the **total** associations between A and B in the universe at large.

As for total association, A and B are said to be **positively associated** in the universe of C's if

$$(ABC) > \frac{(AC)(BC)}{(C)} \qquad . \qquad . \qquad . \qquad (4.1)$$

and **negatively associated** in the converse case.

Similarly they are **positively associated** in the universe of CD's if

$$(ABCD) > \frac{(ACD)(BCD)}{(CD)} \qquad . \qquad . \qquad (4.2)$$

and so on. These formulæ are derived from the formula for total association by specifying the universe in which the partial association exists.

Alternative Forms of the Conditions for Partial Association.

4.5. As in the case of total association, the above forms can be written in many ways, adapted to the nature of the data and of the question which is to be answered. The partial association is most conveniently tested by comparisons of percentages or proportions in the manner of the previous chapter, and we may quote the four most convenient comparisons in the case of three attributes :

$$\frac{(ABC)}{(BC)} > \frac{(AC)}{(C)} \quad . \quad (a) \qquad\qquad \frac{(ABC)}{(AC)} > \frac{(BC)}{(C)} \quad . \quad (b) \left.\begin{array}{c} \\ \\ \\ \end{array}\right\} (4.3)$$

$$\frac{(ABC)}{(BC)} > \frac{(A\beta C)}{(\beta C)} \quad . \quad (c) \qquad\qquad \frac{(ABC)}{(AC)} > \frac{(aBC)}{(aC)} \quad . \quad (d)$$

Similar formulæ may be written down for the cases of four or more attributes, and the methods of this chapter are applicable to such cases. For the sake of simplicity we shall, however, confine ourselves to three attributes hereafter.

4.6. Let us now consider some examples.

Example 4.1.—(Material from ref. (69).)

The following are the proportions per 10,000 of boys observed with certain classes of defects amongst a number of school-children. (A) denotes the number with development defects, (B) the number with nerve signs, (D) the number of the " dull."

N	10,000	(AB)	338
(A)	877	(AD)	338
(B)	1,086	(BD)	455
(D)	789	(ABD)	153

The *Report* from which the figures are drawn concludes that "the connecting link between defects of body and mental dulness is the coincident defect of brain which may be known by observation of abnormal nerve signs." Discuss this conclusion.

The phrase " connecting link " is a little vague, but it may mean that the mental defects indicated by nerve signs B may give rise to development defects A, and also to mental dulness D; A and D being thus common effects of the same cause B (or another attribute necessarily indicated by B) and not directly influencing each other. The case is thus similar to that of the first illustration of **4.2** (liability to smallpox and to non-vaccination being held to be common effects of the same circumstances), and may be similarly treated by investigation of the partial associations between A and D for the universes B and β. As the ratios $(A)/N$, $(B)/N$, $(D)/N$ are small, comparisons of the form (3.5), page 39, or (4.3) (a) and (b) above, may be used.

The following figures illustrate, then, the association between A and D for the whole universe, the B-universe and the β-universe :—

For the entire material :

Proportion of the dull $= (D)/N$ . . . $= \dfrac{789}{10,000} = 7\cdot9$ per cent.

$\qquad$ " $\qquad$ " defectively developed who $\left.\begin{array}{c}\\\end{array}\right\}$ = $\dfrac{338}{877}$ = 38·5 $\qquad$ "
were dull $=(AD)/(A)$ . . . .

For those exhibiting nerve signs :

Proportion of the dull $=(BD)/(B)$ . . $=\dfrac{455}{1,086}$ $=41\cdot9$ per cent.

,, ,, defectively developed who $\bigg\}$ $=\dfrac{153}{338}$ $=45\cdot3$,,
were dull $=(ABD)/(AB)$ . . .

For those not exhibiting nerve signs :

Proportion of the dull $=(\beta D)/(\beta)$ · . . $=\dfrac{334}{8,914}$ $=3\cdot7$,,

,, ,, defectively developed who $\bigg\}$ $=\dfrac{185}{539}$ $=34\cdot3$,,
were dull $=(A\beta D)/(A\beta)$ . . .

The results are extremely striking ; the association between A and D is high both for the material as a whole (the universe at large) and for those not exhibiting nerve signs (the β-universe), but it is *small* for those who do exhibit nerve signs (the B-universe).

This result does not appear to be in accord with the conclusion of the *Report*, as we have interpreted it, for the association between A and D in the β-universe should in that case have been low instead of high.

Example 4.2.—Eye-colour of grandparent, parent and child. (Material from Sir Francis Galton's "*Natural Inheritance*" (1889), Table 20, p. 216. The table only gives particulars for 78 large families with not less than six brothers or sisters, so that the material is hardly entirely representative, but serves as a good illustration of the method.) The original data are treated as in Example 3.7, page 41. Denoting a light-eyed child by A, parent by B, grandparent by C, every possible line of descent is taken into account. Thus, taking the following two lines of the table,

Children.		Parents.		Grandparents.	
A.	a.	B.	β.	C.	γ.
Light-eyed.	Not-Light-eyed.	Light-eyed.	Not-Light-eyed.	Light-eyed.	Not-Light-eyed.
4	5	1	1	1	3
3	4	1	1	4	0

the first would give $4 \times 1 \times 1 = 4$ to the class ABC, $4 \times 1 \times 3 = 12$ to the class $AB\gamma$, 4 to $A\beta C$, 12 to $A\beta\gamma$, 5 to aBC, 15 to $aB\gamma$, 5 to $a\beta C$ and 15 to $a\beta\gamma$; the second would give $3 \times 1 \times 4 = 12$ to the class ABC, 12 to $A\beta C$, 16 to aBC, 16 to $a\beta C$ and none to the remainder. The class-frequencies so derived from the whole table are :

(ABC)	1928	(aBC)	303
$(AB\gamma)$	596	$(aB\gamma)$	225
$(A\beta C)$	552	$(a\beta C)$	395
$(A\beta\gamma)$	508	$(a\beta\gamma)$	501

The following comparisons indicate the association between grandparents and parents, parents and children, and grandparents and grandchildren, respectively :—

Grandparents and Parents.

Proportion of light-eyed amongst the children $\left.\right\}$ of light-eyed grandparents $=\dfrac{(BC)}{(C)}=\dfrac{2231}{3178}=70\cdot2$ per cent.

Proportion of light-eyed amongst the children $\left.\right\}$ of not-light-eyed grandparents $=\dfrac{(B\gamma)}{(\gamma)}=\dfrac{821}{1830}=44\cdot9$,,

Parents and Children.

Proportion of light-eyed amongst the children $\left.\right\}$ of light-eyed parents $=\dfrac{(AB)}{(B)}=\dfrac{2524}{3052}=82\cdot7$ per cent.

Proportion of light-eyed amongst the children $\left.\right\}$ of not-light-eyed parents $=\dfrac{(A\beta)}{(\beta)}=\dfrac{1060}{1956}=54\cdot2$,,

In both the above cases we are really dealing with the association between parent and offspring, and consequently the intensity of association is, as might be expected, approximately the same ; in the next case it is naturally lower :

Grandparents and Grandchildren.

Proportion of light-eyed amongst the grand-$\left.\right\}$ children of light-eyed grandparents $=\dfrac{(AC)}{(C)}=\dfrac{2480}{3178}=78\cdot0$ per cent.

Proportion of light-eyed amongst the grand-$\left.\right\}$ children of not-light-eyed grandparents $=\dfrac{(A\gamma)}{(\gamma)}=\dfrac{1104}{1830}=60\cdot3$,,

We proceed now to test the *partial associations* between grandparents and grandchildren, as distinct from the total associations given above, in order to throw light on the real nature of the resemblance. There are two such partial associations to be tested : (1) where the parents are light-eyed, (2) where they are not-light-eyed. The following are the comparisons :—

Grandparents and Grandchildren : Parents light-eyed.

Proportion of light-eyed amongst the grand-$\left.\right\}$ children of light-eyed grandparents. $=\dfrac{(ABC)}{(BC)}=\dfrac{1928}{2231}=86\cdot4$ per cent.

Proportion of light-eyed amongst the grand-$\left.\right\}$ children of not-light-eyed grandparents $=\dfrac{(AB\gamma)}{(B\gamma)}=\dfrac{596}{821}=72\cdot6$,,

Grandparents and Grandchildren : Parents not-light-eyed.

Proportion of light-eyed amongst the grand-$\left.\right\}$ children of light-eyed grandparents. $=\dfrac{(A\beta C)}{(\beta C)}=\dfrac{552}{947}=58\cdot3$ per cent.

Proportion of light-eyed amongst the grand-$\left.\right\}$ children of not-light-eyed grandparents $=\dfrac{(A\beta\gamma)}{(\beta\gamma)}=\dfrac{508}{1009}=50\cdot3$,,

In both cases the partial association is quite well marked and positive ; the total association between grandparents and grandchildren cannot, then, be due wholly to the total associations between grandparents and parents, parents and children, respectively. There is an *ancestral heredity*, as it is termed, as well as a parental heredity.

We need not discuss the partial association between children and parents, as it is comparatively of little consequence. It may be noted, however, as regards the above results, that the most important feature may be brought out by stating three ratios only.

If A and B are positively associated, $(AB)/(B) > (A)/N$.

If A and C are positively associated in the universe of B's, $(ABC)/(BC) > (AB)/(B)$. Hence $(A)/N$, $(AB)/(B)$ and $(ABC)/(BC)$ form an ascending series. Thus we have from the given data:

Proportion of light-eyed amongst children in general	$= (A)/N$	$=71\cdot6$ per cent.
Proportion of light-eyed amongst the children of light-eyed parents	$= (AB)/(B)$	$=82\cdot7$,,
Proportion of light-eyed amongst the children of light-eyed parents and grandparents .	$=(ABC)/(BC)=86\cdot4$	,,

If the great-grandparents, etc., etc., were also known, the series might be continued, giving $(ABCD)/(BCD)$, $(ABCDE)/(BCDE)$ and so forth. The series would probably ascend continuously though with smaller intervals, A and D being positively associated in the universe of BC's, A and E in the universe of BCD's, etc.

Notation for Partial Associations.

4.7. We now introduce a notation which is analogous to that used for total associations. It will be remembered that in the last chapter we wrote:

$$(AB)_0 = \frac{(A)(B)}{N}$$

$$\delta = (AB) - (AB)_0$$

We now write:

$$(AB \cdot C)_0 = \frac{(AC)(BC)}{(C)}, \quad (AB \cdot CD)_0 = \frac{(ACD)(BCD)}{(CD)} \left.\begin{matrix} \\ \\ \end{matrix}\right\}$$

$$\delta_{AB.C} = (ABC) - (AB \cdot C)_0, \quad \delta_{AB.CD} = (ABCD) - (AB \cdot CD)_0, \text{ etc.} \quad (4.4)$$

The δ-numbers measure the divergence of the actual frequencies from those which would exist if the attributes were independent in the sub-universe under discussion.

It is also possible to generalise the coefficient of association Q by defining partial coefficients of the type

$$Q_{AB.C} = \frac{(ABC)(\alpha\beta C) - (A\beta C)(\alpha BC)}{(ABC)(\alpha\beta C) + (A\beta C)(\alpha BC)} \left.\begin{matrix} \\ \\ \end{matrix}\right.$$
$$= \frac{(C)\delta_{AB.C}}{(ABC)(\alpha\beta C) + (A\beta C)(\alpha BC)} \left.\begin{matrix} \\ \\ \end{matrix}\right\} \quad . \quad (4.5)$$

The student will notice that the formulæ for the δ-numbers and for the Q numbers are obtained from the expressions for total association by specifying the universe in which the partial association is to be considered. They need not therefore be memorised.

Number of Partial Associations.

4.8. For three attributes A, B, C there are three total associations, namely, those of A with B, B with C and C with A; and six partial associations, namely, those of A and B in C and γ, B and C in A and α, and C and A in B and β.

For four attributes there are fifty-four associations ; for we can choose two attributes from four in six ways, and there are nine associations for each pair (one total, four partials in the sub-universes specified by one attribute, and four partials in the sub-universes specified by two).

We state without proof that for n attributes there are $\dfrac{n(n-1)}{2}3^{n-2}$ associations. $\dfrac{n(n-1)}{2}$ of these are total and the remainder partial. For $n > 4$ this number is so large as to be almost unmanageable. For instance, if $n = 5$ it is 270, and if $n = 6$ it is 1215.

4.9. The large number of partial associations which exists might be thought to occasion some difficulty. We may, however, reassure ourselves by two considerations.

In the first place, it is rarely necessary to investigate in any practical instance all the partial associations which are theoretically possible. For instance, in Example 4.1 the total and partial associations between A and D were alone investigated ; those between A and B, B and D were not essential for answering the question which was asked. Again, in Example 4.2 the three total associations and the partial associations between A and C were all that were necessary.

Relations between Partial Associations.

4.10. In the second place, a theoretical discussion of the partial associations is assisted by the following result : The $\dfrac{n(n-1)}{2}3^{n-2}$ associations are all expressible in terms of $2^n - (n+1)$ algebraically independent associations, together with the class-frequencies N, (A), (B), (C), etc.

In fact, we saw in Chapter 1 that all the class-frequencies can be expressed in terms of the positive class-frequencies, which are 2^n in number in the case of n attributes. Hence the frequencies N, (A), (B), (C), etc., of which there are $(n+1)$, together with the $2^n - (n+1)$ other positive frequencies, completely determine the data, and hence determine the associations, which are expressed in terms of the data. Hence the number of algebraically independent associations which can be derived is only $2^n - (n+1)$.

4.11. In practice the existence of these relations is of little or no value. The formal relations between the ratios and the δ-numbers which express the associations are, in fact, so complex that lengthy algebraic manipulation is necessary to express those which are not known in terms of those which are. It is usually better to evaluate the class-frequencies and calculate the desired results directly from them.

4.12. There is, however, one result which has important theoretical consequences.

We have, by definition,

$$\delta_{AB.C} = (ABC) - \frac{(AC)(BC)}{(C)}$$

$$\delta_{AB.\gamma} = (AB\gamma) - \frac{(A\gamma)(B\gamma)}{(\gamma)}$$

Hence,

$$\delta_{AB.C} + \delta_{AB.\gamma} = (AB) - \frac{1}{(C)(\gamma)}\{(AC)(BC)(\gamma) + (A\gamma)(B\gamma)(C)\}$$

$$= (AB) - \frac{1}{(C)(\gamma)}\{N(AC)(BC) - (A)(C)(BC) - (B)(C)(AC)$$
$$+ (A)(B)(C)\}$$

$$= (AB) - \frac{(A)(B)}{N} - \frac{N}{(C)(\gamma)}\left\{(AC) - \frac{(A)(C)}{N}\right\}\left\{(BC) - \frac{(B)(C)}{N}\right\}$$

$$= \delta_{AB} - \frac{N}{(C)(\gamma)}\delta_{AC}\delta_{BC} \qquad . \qquad . \qquad . \qquad . \qquad . \qquad . \qquad (4.6)$$

This gives us the sum of the δ-numbers for the partial associations of A and B in C and γ in terms of the total associations between A, B and C.

Now suppose that A and B are independent in C and γ. Then we have :

$$\delta_{AB.C} = \delta_{AB.\gamma} = 0$$

and

$$\delta_{AB} = \frac{N}{(C)(\gamma)}\delta_{AC}\delta_{BC}$$

δ_{AB} is not zero unless one or both of δ_{AC}, δ_{BC} are zero.

Hence, if A and B are independent within the universes of C's and not-C's, they will nevertheless be associated in the universe at large unless C is independent of A or B or both.

Illusory Associations.

4.13. This peculiar result indicates that, although a set of attributes independent of A and B will not affect the association between them, the existence of an attribute C with which they are both associated may give an association in the universe at large which is illusory in the sense that it does not correspond to any real relationship between them. If the associations between A and C, B and C are of the same sign, the resulting association between A and B will be positive ; if of opposite signs, negative.

The cases which we discussed at the beginning of this chapter are instances in point. In the first illustration we saw that it was possible to argue that the positive associations between *vaccination* and *hygienic conditions*, *exemption from attack* and *hygienic conditions*, led to an illusory association between *vaccination* and *exemption from attack*. Similarly, the question was raised whether the positive association between *grandfather* and *grandchild* may not be due to the positive associations between *grandfather* and *father*, and *father* and *child*.

4.14. Misleading associations may easily arise through the mingling of records which a careful worker would keep distinct.

Take the following case, for example. Suppose there have been 200 patients in a hospital, 100 males and 100 females, suffering from some disease. Suppose, further, that the death-rate for males (the case mortality) has been 30 per cent., for females 60 per cent. A new treatment is tried on 80 per cent. of the males and 40 per cent. of the females, and the results published without distinction of sex. The three attributes, with

the relations of which we are here concerned, are *death, treatment* and *male sex.* The data show that more males were treated than females, and more females died than males; therefore the first attribute is associated negatively, the second positively, with the third. It follows that there will be an illusory negative association between the first two—*death* and *treatment.* If the treatment were completely inefficient we should, in fact, have the following results :—

	Males.	Females.	Total.
Treated and died . . .	24	24	48
,, and did not die .	56	16	72
Not treated and died . .	6	36	42
,, and did not die .	14	24	38

i.e. of the treated, only $48/120 = 40$ per cent. died, while of those not treated $42/80 = 52.5$ per cent. died. If this result were stated without any reference to the fact of the mixture of the sexes, to the different proportions of the two that were treated and to the different death-rates under normal treatment, then some value in the new treatment would appear to be suggested. To make a fair return, either the results for the two sexes should be stated separately, or the same proportion of the two sexes must receive the experimental treatment. Further, care would have to be taken in such a case to see that there was no selection (perhaps unconscious) of the less severe cases for treatment, thus introducing another source of fallacy (*death* positively associated with *severity, treatment* negatively associated with *severity,* giving rise to illusory negative association between *treatment* and *death*).

4.15. Illusory associations may also arise in a different way through the personality of the observer or observers. If the observer's attention fluctuates, he may be more likely to notice the presence of A when he notices the presence of B, and *vice versa* ; in such a case A and B (so far as the record goes) will both be associated with the observer's attention C, and consequently an illusory association will be created. Again, if the attributes are not well defined, one observer may be more generous than another in deciding when to record the presence of A and also the presence of B, and even one observer may fluctuate in the generosity of his marking. In this case the recording of A and the recording of B will both be associated with the generosity of the observer in recording their presence, C, and an illusory association between A and B will consequently arise, as before.

Determination of Sign of Association when the Data are Incomplete.

4.16. It is important to notice that, though we cannot actually determine the partial associations unless the third-order frequency (ABC) is given, we can make some conjecture as to their signs from the values of the second-order frequencies.

In **4.12** we have:

$$\delta_{AB.C} + \delta_{AB.\gamma} = (AB) - \frac{(AC)(BC)}{(C)} - \frac{(A\gamma)(B\gamma)}{(\gamma)} . \qquad . \quad (4.7)$$

Hence, if the expression on the right is positive, one at least of $\delta_{AB.c}$, $\delta_{AB.\gamma}$ is positive, *i.e.* A and B are positively associated either in C or γ or both. Similarly, if the expression is negative, A and B are negatively

associated either in C or in γ or in both. Finally, if the expression is zero, A and B are either independent in both C and γ, or positively associated in one and negatively in the other.

The expression (4.7) may be thrown into a form more convenient when percentages are given. Dividing through by (B) we have:

$$\frac{\delta_{AB.C} + \delta_{AB.\gamma}}{(B)} = \frac{(AB)}{(B)} - \frac{(AC)}{(C)}\frac{(BC)}{(B)} - \frac{(A\gamma)}{(\gamma)}\frac{(B\gamma)}{(B)} \qquad . \quad (4.8)$$

The following examples illustrate the method.

Example 4.3.—(Figures compiled from *Supplement to the Fifty-fifth Annual Report of the Registrar-General* [C.—8503], 1897.) The following are the death-rates per thousand per annum, and the proportions over 65 years of age, of occupied males in general, farmers, textile workers and glass workers (over 15 years of age in each case) during the decade 1891–1900 in England and Wales.

	Death-rate per thousand.	Proportion per thousand over 65 Years of Age.
Occupied males over 15 . .	15·8	46
Farmers, ,, ,, . .	19·6	132
Textile workers, males over 15	15·9	34
Glass workers, ,, .,	16·6	16

Would farming, textile working and glass working seem to be relatively healthy or unhealthy occupations, given that the death-rates among occupied males from 15–65 and over 65 years of age are 11·5 and 102·3 per thousand, respectively?

If A denote *death*, B the given *occupation*, C *old age*, we have to apply the principle of equation (4.8), calculate what would be the death-rate for each occupation on the supposition that the death-rates for occupied males in general (11·5, 102·3) apply to each of its separate age-groups (under 65, over 65), and see whether the total death-rate so calculated exceeds or falls short of the actual death-rate. If it exceeds the actual rate, the occupation must on the whole be healthy; it if falls short, unhealthy. Thus we have the following calculated death-rates :—

Farmers . . .	$11·5 \times 0·868 + 102·3 \times 0·132 = 23·5$
Textile workers .	$11·5 \times 0·966 + 102·3 \times 0·034 = 14·6$
Glass workers . .	$11·5 \times 0·984 + 102·3 \times 0·016 = 13·0$

The calculated rate for farmers largely exceeds the actual rate; farming then must, on the whole, as one would expect, be a healthy occupation. The death-rate for either young farmers or old farmers, or both, must be less than for occupied males in general (the last is actually the case); the high death-rate observed is due solely to the large proportion of the aged. Textile working, on the other hand, appears to be unhealthy (14·6 < 15·9), and glass working still more so (13·0 < 16·6); the actual low total death-rates are due merely to low proportions of the aged.

It is evident that age-distributions vary so largely from one occupation to another that total death-rates are liable to be very misleading—so misleading, in fact, that they are not tabulated at all by the Registrar-General ; only death-rates for narrow limits of age (5 or 10 year age-classes) are worked out. Similar fallacies are liable to occur in comparisons of local death-rates, owing to variations not only in the relative proportions of the old, but also in the relative proportions of the two sexes.

It is hardly necessary to observe that as *age* is a variable quantity, the above procedure for calculating the comparative death-rates is extremely rough. The death-rate of those engaged in any occupation depends not only on the mere proportions over and under 65, but on the relative numbers at every single year of age. The simpler procedure brings out, however, better than a more complex one, the nature of the fallacy involved in assuming that crude death-rates are measures of healthiness.

Example 4.4.—Eye-colour in grandparent, parent and child. (The figures are those of Example 4.2.)

A, light-eyed child ; B, light-eyed parent ; C, light-eyed grandparent.

$$\begin{aligned} N &= 5008 & (AB) &= 2524 \\ (A) &= 3584 & (AC) &= 2480 \\ (B) &= 3052 & (BC) &= 2231 \\ (C) &= 3178 \end{aligned}$$

Given only the above data, investigate whether there is probably a partial association between child and grandparent.

If there were no partial association we should have :

$$\begin{aligned} (AC) &= \frac{(AB)(BC)}{(B)} + \frac{(A\beta)(\beta C)}{(\beta)} \\ &= \frac{2524 \times 2231}{3052} + \frac{1060 \times 947}{1956} \\ &= 1845 \cdot 0 + 513 \cdot 2 \\ &= 2358 \cdot 2 \end{aligned}$$

Actually $(AC) = 2480$; there must, then, be partial association either in the B-universe, the β-universe, or both. In the absence of any reason to the contrary, it would be natural to suppose there is a partial association in both, *i.e.* that there is a partial association with the grandparent whether the line of descent passes through " light-eyed " or " not-light-eyed " parents ; but this could not be *proved* without a knowledge of the class-frequency (ABC).

Complete Independence.

4.17. The particular case in which all the $2^n - (n + 1)$ given associations are zero is worth some special investigation.

It follows, in the first place, that all other possible associations must be zero, *i.e.* that a state of **complete independence,** as we may term it, exists. Suppose, for instance, that we are given :

$$(AB) = \frac{(A)(B)}{N} \qquad (AC) = \frac{(A)(C)}{N}$$

$$(BC) = \frac{(B)(C)}{N} \qquad (ABC) = \frac{(AC)(BC)}{(C)} = \frac{(A)(B)(C)}{N^2}$$

Then it follows at once that we have also:

$$(ABC) = \frac{(AB)(BC)}{(B)} = \frac{(AB)(AC)}{(A)}$$

i.e. A and C are independent in the universe of B's, and B and C in the universe of A's. Again,

$$(AB\gamma) = (AB) - (ABC) = \frac{(A)(B)}{N} - \frac{(A)(B)(C)}{N^2}$$

$$= \frac{(A)(B)(\gamma)}{N^2} = \frac{(A\gamma)(B\gamma)}{(\gamma)}$$

Therefore A and B are independent in the universe of γ's. Similarly, it may be shown that A and C are independent in the universe of β's, B and C in the universe of α's.

In the next place it is evident from the above that relations of the general form (to write the equation symmetrically)

$$\frac{(ABC)}{N} = \frac{(A)}{N} \cdot \frac{(B)}{N} \cdot \frac{(C)}{N} \qquad . \qquad . \qquad . \quad (4.9)$$

must hold for every class-frequency. This relation is the general form of the equation of independence (3.2) (*d*), page 35.

4.18. It must be noted, however, that (4.9) is not a *criterion* for the *complete independence* of A, B and C in the sense that the equation

$$\frac{(AB)}{N} = \frac{(A)}{N} \cdot \frac{(B)}{N}$$

is a criterion for the complete independence of A and B. If we are given N, (A) and (B), and the last relation quoted holds good, we know that similar relations must hold for (Aβ), (αB) and (αβ). If N, (A), (B) and (C) be given, however, and the equation (4.9) holds good, we can draw no conclusion without further information; the data are insufficient. There are *eight* algebraically independent class-frequencies in the case of three attributes, while N, (A), (B), (C) are only four: the equation (4.9) must therefore be shown to hold good for *four* frequencies of the third order before the conclusion can be drawn that it holds good for the remainder, *i.e.* that a state of complete independence subsists. The direct verification of this result is left for the student.

Quite generally, if N, (A), (B), (C), . . . be given, the relation

$$\frac{(ABC \ldots)}{N} = \frac{(A)}{N} \cdot \frac{(B)}{N} \cdot \frac{(C)}{N} \ldots \qquad . \qquad . \quad (4.10)$$

must be shown to hold good for $2^n - (n+1)$ of the nth order classes before it may be assumed to hold good for the remainder. It is only because

$$2^n - (n+1) = 1$$

when $n = 2$ that the relation

$$\frac{(AB)}{N} = \frac{(A)}{N} \cdot \frac{(B)}{N}$$

may be treated as a *criterion* for the independence of A and B. *If* all the n $(n > 2)$ attributes are completely independent, the relation (4.10) holds good; but it does not follow that if the relation (4.10) holds good they are all independent.

SUMMARY.

1. The association of A and B in sub-universes of the type C, $\dot{\gamma}$, CD, CDE, etc. is called a partial association.

2. If

$$(ABC) > \frac{(AC)(BC)}{(C)}$$

A and B are positively associated in C; and if

$$(ABC) < \frac{(AC)(BC)}{(C)}$$

A and B are negatively associated in C.

3. There are $\dfrac{n(n-1)}{2} 3^{n-2}$ associations in a universe characterised by n attributes, $\dfrac{n(n-1)}{2}$ of which are total and the remainder partial.

4. All the associations are expressible in terms of N, (A), (B), (C), etc., and $2^n - (n+1)$ algebraically independent associations. These relations have, however, only a theoretical value.

5. If A and B are independent within the universe of C's they will nevertheless be associated within the universe at large, unless C is independent of either A or B or both.

6. In interpreting an association between A and B it must be remembered that this may arise owing to associations of A with C and B with C. To resolve this point it is necessary to consider the partial associations of A and B in C and γ.

7. Complete independence of n attributes occurs if $2^n - (n+1)$ algebraically independent associations and hence all associations are zero. In this case

$$\frac{(ABC \ \ldots)}{N} = \frac{(A)}{N} \ \frac{(B)}{N} \ \frac{(C)}{N} \ \cdots$$

but this last condition is not sufficient for complete independence.

EXERCISES.

4.1. Take the following figures for girls corresponding to those for boys in Example 4.1, page 52, and discuss them similarly, but not necessarily using exactly the same comparisons, to see whether the conclusion that "the connecting link between defects of body and mental dulness is the coincident defect of brain which may be known by observation of abnormal nerve signs" seems to hold good.

A, development defects; B, nerve signs; D, mental dulness.

N	10,000	(AB)	248
(A)	682	(AD)	307
(B)	850	(BD)	363
(D)	689	(ABD)	128

4.2. (Material from *Census of England and Wales, 1891*, vol. 3.) The following figures give the numbers of those suffering from single or combined infirmities: (1) for all males; (2) for males of 55 years of age and over.

A, blindness; B, mental derangement; C, deaf-mutism.

	(1)	(2)		(1)	(2)
	All Males.	Males 55–.		All Males.	Males 55–.
N	14,053,000	1,377,000	(AB)	183	65
(A)	12,281	5,538	(AC)	51	14
(B)	45,392	10,309	(BC)	299	47
(C)	7,707	746	(ABC)	11	3

Tabulate proportions per thousand, exhibiting the total association between blindness and mental derangement, and the partial association between the same two infirmities among deaf-mutes: (1) for males in general; (2) for those of 55 years of age and over. Give a short verbal statement of the results, and contrast them with those of Exercise 4.1.

4.3. (Material from *Supplement to Fifty-fifth Annual Report of the Registrar-General.*)

The death-rate from cancer for occupied males in general (over 15) is 0·685 per thousand per annum, and for farmers 1·20.

The death-rates from cancer for occupied males under and over 45 respectively are 0·13 and 2·25 respectively. Of the farmers, 46·1 per cent. are over 45.

Would you say that farmers were peculiarly liable to cancer?

4.4. A population of males over 15 years of age consists of 7 per cent. over 65 years of age and 93 per cent. under. The death-rates are 12 per thousand per annum in the younger class and 110 in the older, or 18·86 in the whole population. The death-rate of males (over 15) engaged in a certain industry is 26·7 per thousand.

If the industry be not unhealthy, what must be the approximate proportion of those over 65 engaged in it (neglecting minor differences of age distribution)?

4.5. Show that if A and B are independent, while A and C, B and C are associated, A and B must be disassociated either in the universe of C's, the universe of γ's, or both.

4.6. As an illustration of Exercise 4.5, show that if the following were actual data, there would be a slight disassociation between the eye-colours of husband and wife (father and mother) for the parents either of light-eyed sons or not-light-eyed sons, or both, although there is a slight positive association for parents at large.

A light eye-colour in husband, B in wife, C in son :

N	1000	(AB)	358
(A)	622	(AC)	471
(B)	558	(BC)	419
(C)	617		

4.7. Show that if $(ABC)=(a\beta\gamma)$, $(aBC)=(A\beta\gamma)$, and so on (the case of "complete equality of contrary frequencies" of Exercise 1.7, page 23), A, B and C are completely independent if A and B, A and C, B and C are independent pair and pair.

4.8. If, in the same case of complete equality of contraries,

$$(AB) - N/4 = \delta_1$$
$$(AC) - N/4 = \delta_2$$
$$(BC) - N/4 = \delta_3$$

show that

$$2\left[(ABC) - \frac{(AC)(BC)}{(C)}\right] = 2\left[(AB\gamma) - \frac{(A\gamma)(B\gamma)}{(\gamma)}\right] = \delta_1 - \frac{4\delta_2\delta_3}{N}$$

so that the partial associations between A and B in the universes C and γ are positive or negative according as

$$\delta_1 \gtrless \frac{4\delta_2\delta_3}{N}$$

4.9. In the simple contests of a general election (contests in which one Conservative opposed one Socialist and there were no other candidates) 66 per cent. of the winning candidates (according to the returns) spent more money than their opponents. Given that 63 per cent. of the winners were Conservatives, and that the Conservative expenditure exceeded the Socialist in 80 per cent. of the contests, find the percentages of elections won by Conservatives (1) when they spent more and (2) when they spent less than their opponents, and hence say whether you consider the above figures evidence of the influence of expenditure on election results or no. (*Note* that if the one candidate in a contest be a *Conservative-winner-who spends more than his opponent*, the other must necessarily be a *Socialist-loser-who spends less*—and so forth. Hence the case is one of complete equality of contraries.)

4.10. Given that $(A)/N = (B)/N = (C)/N = x$, and that $(AB)/N = (AC)/N = y$, find the major and minor limits to y that enable one to infer positive association between B and C, *i.e.* $(BC)/N > x^2$.

Draw a diagram on squared paper to illustrate your answer, taking x and y as co-ordinates, and shading the limits within which y must lie in order to permit of the above inference. Point out the peculiarities in the case of inferring a positive association from two negative associations.

4.11. Discuss similarly the more complex case $(A)/N = x$, $(B)/N = 2x$, $(C)/N = 3x$:

(1) for inferring positive association between B and C given $(AB)/N = (AC)/N = y$.

(2) for inferring positive association between A and C given $(AB)/N = (BC)/N = y$.

(3) for inferring positive association between A and B given $(AC)/N = (BC)/N = y$.

CHAPTER 5.

MANIFOLD CLASSIFICATION.

Manifold Classification.

5.1. Instead of dividing the universe of discourse into two parts by a simple dichotomy, we may also divide it into a number of parts by a similar process. For instance, we can extend the dichotomy of the universe of men into " those with blue eyes " and " those not with blue eyes " to a threefold division: "those with blue eyes," "those with brown eyes," and " those with neither blue nor brown eyes "; or into a fourfold division by adding a fresh category, "those with grey eyes"; and so on.

Generally, our universe may be divided first according to s heads, A_1, A_2, . . . A_s; each of the classes so obtained into t heads, B_1, B_2, . . . B_t; each of these into u heads, C_1, C_2, . . . C_u; and so on.

This is called **manifold classification.**

5.2. The general theory of manifold classification for n attributes is rather complicated, but its fundamental principles are very similar to those which apply to dichotomy. A straightforward extension of the methods of Chapter 1 will give the following results, which we are content to announce without a formal proof :—

   (*a*) There are $s \times t \times u \times$ . . . ultimate classes.

   (*b*) The total number of classes, including N and the ultimate classes, is $(s+1)(t+1)(u+1)$ . . .

   (*c*) The data are consistent if, and only if, every ultimate class-frequency is not negative.

   (*d*) The data are completely specified by $s \times t \times u \times$ . . . algebraically independent class-frequencies. Even if all these are not given, it may be possible to set limits to the other class-frequencies.

For example, if the population of the United Kingdom is classified geographically according to habitation in England, Wales, Scotland and Northern Ireland ; by eye-colour into blue, brown, grey, green and the remainder ; and by hair-colour into black, fair, red and the remainder ; there will be 150 classes altogether, expressible in terms of 80 independent class-frequencies.

5.3. Data so completely specified are very rare, and an elaborate discussion of the general case would hardly be justified by its practical value. For the remainder of this chapter, therefore, we shall be concerned solely with the case of two characteristics, A and B.

Contingency Tables.

5.4. Let us suppose that the classification of the A's is s-fold and that of the B's is t-fold. Then there will be st classes of the type $A_m B_n$.

Generalising slightly the notation of previous chapters, let the frequency of individuals A_m be denoted by (A_m) and of individuals A_mB_n by (A_mB_n). The data can then be set out in the form of a table of t rows and s columns as-follows :—

TABLE 5.1.

Attribute A

	A_1	A_2	—	—	A_{s-1}	A_s	Totals.
B_1	(A_1B_1)	(A_2B_1)	—	—	$(A_{s-1}B_1)$	(A_sB_1)	(B_1)
B_2	(A_1B_2)	(A_2B_2)	··	—	$(A_{s-1}B_2)$	(A_sB_2)	(B_2)
—	—	—	—·	—	—	—·	—
B_t	(A_1B_t)	(A_2B_t)	—	—	$(A_{s-1}B_t)$	(A_sB_t)	(B_t)
Totals	(A_1)	(A_2)	—	—	(A_{s-1})	(A_s)	N

Attribute B.

In this table the frequency of the class A_mB_n is entered in the compartment common to the mth column and the nth row ; the totals at the ends of rows and at the feet of columns give the first order frequencies, *i.e.* the numbers of A_m's and B_n's ; and finally, the grand total in the bottom right-hand corner gives the whole number of observations.

Such a table is called a **contingency table**. It is a generalised form of the fourfold (2×2-fold) table in **3.1**.

Example 5.1.—In Table 5.2 below the classification is 3×4-fold : the eye-colours are classed under the three heads " blue," " grey or green " and " brown," while the hair-colours are classed under four heads, " fair," " brown," " black " and " red." Taking the first row,

TABLE 5.2.—*Hair- and Eye-colours of 6800 Males in Baden.*
(Ammon, *Zur Anthropologie der Badener.*)

Eye-colour.	Hair-colour.				Total.
	Fair.	Brown.	Black.	Red.	
Blue	1768	807	189	47	2811
Grey or Green . .	946	1387	746	53	3132
Brown . .	115	438	288	16	857
Total .	2829	2632	1223	116	6800

the table tells us that there were 2811 men with blue eyes noted, of whom 1768 had fair hair, 807 brown hair, 189 black hair and 47 red hair. Similarly, from the first column, there were 2829 men with fair hair, of whom 1768 had blue eyes, 946 grey or green eyes and 115 brown eyes.

Association in Contingency Tables.

5.5. For the purpose of discussing the nature of the relation between the A's and the B's, any such table may be treated on the principles of the preceding chapters by reducing it in different ways to a 2×2-fold form. It then becomes possible to trace the association between any one or more of the A's and any one or more of the B's, either in the universe at large or in universes limited by the omission of one or more of the A's, of the B's, or of both.

If, *e.g.*, we desire to trace the association between a lack of pigmentation in eyes and in hair, rows 1 and 2 may be pooled together as representing the least pigmentation of the eyes, and columns 2, 3 and 4 may be pooled together as representing hair with a more or less marked degree of pigmentation. We then have:

Proportion of light-eyed with fair hair . . . . $2714/5943 = 46$ per cent.

Proportion of brown-eyed with fair hair . . . . $115/857 \ \ = 13$,,

The association is therefore well marked. For comparison we may trace the corresponding association between the most marked degree of pigmentation in eyes and hair, *i.e.* brown eyes and black hair. Here we must add together rows 1 and 2 as before, and pool columns 1, 2 and 4 —the column for red being really misplaced, as red represents a comparatively slight degree of pigmentation. The figures are:

Proportion of brown-eyed with black hair . . . $288/857 \ \ = 34$ per cent.

Proportion of light-eyed with black hair . . . $935/5943 = 16$,,

The association is again positive and well marked, but the difference between the two percentages is rather less than in the last case.

5.6. The mode of treatment adopted in the preceding two paragraphs rests on first principles and, if fully carried out, gives us all the information possible about the associations of the two attributes. At the same time, it is laborious if s and t are at all large. Moreover, in practical work we are often concerned, not with the associations of individual A's with individual B's, but with finding the answer to a general question of the type : Are the A's *on the whole* distinctly dependent on the B's, and if so, is this dependence very close, or the reverse ? In fact, what we want is a coefficient which will summarise the general nature of the dependence. We will proceed to discuss two such coefficients.

Coefficients of Contingency.

5.7. If the A's and B's be completely independent in the universe at large, we must have for all values of m and n:

$$(A_m B_n) = \frac{(A_m)(B_n)}{N} = (A_m B_n)_0 \quad . \qquad . \qquad . \quad (5.1)$$

If, however, A and B are not completely independent, $(A_m B_n)$ and $(A_m B_n)_0$

will not be identical for all values of m and n. Let the difference be given by

$$\delta_{mn} = (A_m B_n) - (A_m B_n)_0 \qquad . \qquad . \qquad . \quad (5.2)$$

Let us note in passing the following properties of these quantities:

(1) In the first place, δ_{mn} is not equal to δ_{nm}.

(2) In the second place, the δ's are not all algebraically independent.

We have, in fact, for any particular m:

$$\delta_{m1} + \delta_{m2} + \delta_{m3} + \ldots + \delta_{mn} + \ldots + \delta_{mt}$$

$$= (A_m B_1) - \frac{(A_m)(B_1)}{N} + (A_m B_2) - \frac{(A_m)(B_2)}{N} \ldots + (A_m B_t) - \frac{(A_m)(B_t)}{N}$$

$$= (A_m) - \frac{(A_m)}{N}\{(B_1) + (B_2) + \ldots + (B_t)\}$$

$$= 0 \quad . \qquad . \qquad . \qquad . \qquad . \qquad . \qquad . \qquad . \qquad . \qquad . \quad (5.3)$$

A similar relation is true for any particular n.

Now there are st δ-quantities. In virtue of the relationship we have just proved, for any particular m only $(t-1)$ of the t-quantities δ_{mn} are independent. Similarly, for any n only $(s-1)$ are independent. Hence the total number of independent δ's is $(s-1)(t-1)$.

5.8. These δ-quantities indicate the extent of the associations, and we expect a summarising coefficient to be built up from them in some way. It would, however, be useless to add them together, for in virtue of the relation of the preceding paragraph the sum is zero. We wish to construct a coefficient which shall be independent of the signs of the δ-numbers.

We therefore define

$$\chi^2 = S\left(\frac{\delta^2_{mn}}{(A_m B_n)_0}\right) \qquad . \qquad . \qquad . \quad (5.4)$$

and call χ^2 the " square contingency."

We then write:

$$\phi^2 = \frac{\chi^2}{N} \qquad . \qquad . \qquad . \qquad . \quad (5.5)$$

and call ϕ^2 the " mean square contingency."

Clearly χ^2 and ϕ^2, being the sums of squares, cannot be negative. They vanish if, and only if, every δ-number vanishes, in which case A and B are independent.

Pearson's Coefficient of Mean Square Contingency.

5.9. The quantity ϕ^2 is not quite suitable in itself to form a coefficient, because its limits vary in different cases. Karl Pearson therefore proposed the coefficient C, defined by

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} = \sqrt{\frac{\phi^2}{1 + \phi^2}} \qquad . \qquad . \quad (5.6)$$

This is called the **Coefficient of Mean Square Contingency**. In general, no sign should be attached to the root, for the coefficient merely shows whether two characters are or are not independent; but in certain cases a conventional sign may be used. Thus, in Table 5.2 slight pigmentation

of eyes and hair appear to go together, and the contingency may be regarded as positive. If *slight* pigmentation of eyes had been associated with *marked* pigmentation of hair, the contingency might have been regarded as negative.

5.10. The coefficient C has one serious disadvantage. Although, as may be seen from its definition, it increases with ϕ^2 towards a limit 1, it never reaches that limit. In fact, the maximum value which it can attain depends on s and t, and reaches unity only for an infinite number of classes. This may be briefly illustrated as follows. Replacing δ_{mn} in equation (5.4) by its value in terms of (A_mB_n) and $(A_mB_n)_0$, we have:

$$\chi^2 = S\left\{\frac{(A_mB_n)^2}{(A_mB_n)_0}\right\} - N \qquad . \qquad . \qquad . \qquad (5.7)$$

and therefore, denoting the summation by S,

$$C = \sqrt{\frac{S-N}{S}} \qquad . \qquad . \qquad . \qquad (5.8)$$

Now suppose we have to deal with a $t \times t$-fold classification in which $(A_m) = (B_m)$ for all values of m; and suppose, further, that the association between A_m and B_m is perfect, so that $(A_mB_m) = (A_m) = (B_m)$ for all values of m, the remaining frequencies of the second order being zero; all the frequency is then concentrated in the diagonal compartments of the table, and each contributes N to the summation S. The total value of S is accordingly tN, and the value of C:

$$C = \sqrt{\frac{t-1}{t}}$$

This is the greatest possible value of C for a symmetrical $t \times t$-fold classification, and therefore, in such a table, for:

$t = 2,$	C cannot exceed		0·707
$t = 3$	,,	,,	0·816
$t = 4$	,,	,,	0·866
$t = 5$	,,	,,	0·894
$t = 6$	,,	,,	0·913
$t = 7$	,,	,,	0·926
$t = 8$	,,	,,	0·935
$t = 9$	,,	,,	0·943
$t = 10$	,,	,,	0·949

5.11. Hence, coefficients calculated from different systems of classification are not, strictly speaking, comparable. This is clearly undesirable. Two coefficients calculated from the same data classified in two different groupings ought not to be very different.

It is as well, therefore, to restrict the use of the C-coefficient to 5×5 or finer groupings. At the same time, the classification must not be made too fine, or the value of the coefficient is largely affected by casual irregularities arising from sampling fluctuations.[1]

[1] Karl Pearson (ref. (86) and in several other papers) has discussed a "correction" to be made to C calculated from coarsely grouped data. The use of such corrections depends to some extent on assumptions about the universe, and may be regarded as attempts to bring the value of C closer to a putative coefficient of correlation (*cf.* **12.20**).

Tschuprow's Coefficient.

5.12. To remedy the defect to which we have just referred, Tschuprow has proposed the coefficient T, defined by

$$T^2 = \frac{\phi^2}{\sqrt{(s-1)(t-1)}} \qquad . \qquad . \qquad . \qquad (5.9)$$

This coefficient varies between 0 and 1 in the desired manner when $s = t$.

We have

$$C^2 = \frac{\phi^2}{1+\phi^2}$$

$$= \frac{\sqrt{(s-1)(t-1)}T^2}{1+\sqrt{(s-1)(t-1)}T^2} \qquad . \qquad . \qquad . \qquad (5.10)$$

and conversely,

$$T^2 = \frac{C^2}{(1-C^2)\sqrt{(s-1)(t-1)}} \qquad . \qquad . \qquad . \qquad (5.11)$$

Calculation of C and T.

5.13. The calculation of C and T is simplified by the use of equation (5.8), which enables us to replace the calculation of the δ's by calculations based on frequencies of types (A_m), (B_n) and $(A_m B_n)$. All these quantities are contained in the contingency tables. The following example will illustrate the method:

Example 5.2.—Consider the data of Table 5.2. (The classification is only 3×4-fold and is therefore rather crude for calculating C, but it will serve as an illustration of the form of the arithmetic.)

We require first of all the quantities $(A_m B_n)_0$, *i.e.* the " independence " values. These are calculated directly from their definition

$$(A_m B_n)_0 = \frac{(A_m)(B_n)}{N}$$

and thus the value for the compartment in the mth column and nth row is the product of the total frequencies in that column and row divided by the whole frequency, *e.g.* $(A_1 B_1)_0 = 2829 \times 2811/6800 = 1169$, and so on.

It is convenient to tabulate the frequencies so obtained in a second contingency table, as in Table 5.3.

TABLE 5.3.—*Independence Values of the Frequencies for Table 5.2.*

Eye colour.	Fair.	Brown.	Black.	Red.
Blue	1169	1088	506	48·0
Grey or Green	1303	1212	563	53·4
Brown	357	332	154	14·6

We now calculate the quantities $\dfrac{(A_m B_n)^2}{(A_m B_n)_0}$.

$(1768)^2/1169$	$2673\cdot9$
$(946)^2/1303$	$686\cdot8$
$(115)^2/357$	$37\cdot0$
$(807)^2/1088$	$598\cdot6$
$(1387)^2/1212$	$1587\cdot3$
$(438)^2/332$	$577\cdot8$
$(189)^2/506$	$70\cdot6$
$(746)^2/563$	$988\cdot5$
$(288)^2/154$	$538\cdot6$
$(47)^2/48\cdot0$	$46\cdot0$
$(53)^2/53\cdot4$	$52\cdot6$
$(16)^2/14\cdot6$	$17\cdot5$
Total $= S =$	$7875\cdot2$

From equation (5.8):

$$C = \sqrt{\frac{S-N}{S}} = \sqrt{\frac{1075\cdot2}{7875\cdot2}}$$

$$= \sqrt{0\cdot1365} = 0\cdot37$$

and

$$T^2 = \frac{C^2}{(1-C^2)\sqrt{(s-1)(t-1)}}$$

$$= \frac{0\cdot1365}{0\cdot8635\sqrt{6}}$$

$$T = \sqrt{0\cdot0645}$$

$$= 0\cdot25$$

The squares in such work may conveniently be taken from Barlow's "*Tables of Squares, Cubes, etc.*," or logarithms may be used throughout— five-figure logarithms are quite sufficient.

It will be seen that T is less than C. This is not always true. Whichever coefficient we use, however, the contingency between pigmentation of hair and eye is evident.

5.14. While such coefficients of contingency are a great convenience in many forms of work, their use should not lead to a neglect of the more detailed treatment of **5.5.** Whether the coefficients be calculated or no, every table should always be examined with care to see if it exhibits any apparently significant peculiarities in the distribution of frequency, *e.g.* in the associations subsisting between A_m and B_n in limited universes. A good deal of caution must be used in order not to be misled by casual irregularities due to paucity of observations in some compartments of the table, but important points that would otherwise be overlooked will often be revealed by such a detailed examination.

5.15. Suppose, for example, that any four adjacent frequencies, say

$$\begin{array}{ll} (A_m B_n) & (A_{m+1} B_n) \\ (A_m B_{n+1}) & (A_{m+1} B_{n+1}) \end{array}$$

are extracted from the general contingency table. If these are considered as a table exhibiting the association between A_m and B_n in a universe limited to A_m A_{m+1} B_n B_{n+1} alone, the association is positive, negative or zero according as $(A_m B_n)/(A_{m+1}B_n)$ is greater than, less than, or equal to the ratio $(A_m B_{n+1})/(A_{m+1}B_{n+1})$. The whole of the contingency table can be analysed into a series of elementary groups of four frequencies like the above, each one overlapping its neighbours, so that an $s \times t$-fold table contains $(s-1)(t-1)$ such "tetrads," and the associations in them all can be very quickly determined by simply tabulating the ratios like $(A_m B_n)/(A_{m+1}B_n)$, $(A_m B_{n+1})/(A_{m+1}B_{n+1})$, etc., or perhaps better, the proportions $(A_m B_n)/\{(A_m B_n) + (A_{m+1}B_n)\}$, etc., for every pair of columns or of rows, as may be most convenient. Taking the figures of Table 5.2 as an illustration, and working from the rows, the proportions run as follows :—

For rows 1 and 2.		For rows 2 and 3.	
1768/2714	0·651	946/1061	0·892
807/2194	0·368	1387/1825	0·760
189/935	0·202	746/1034	0·721
47/100	0·470	53/69	0·768

In both cases the first three ratios form descending series, but the fourth ratio is greater than the second. The signs of the associations in the six tetrads are, accordingly,

$$+ \qquad + \qquad -$$
$$+ \qquad + \qquad -$$

The negative sign in the two tetrads on the right is striking, the more so as other tables for hair- and eye-colour, arranged in the same way, exhibit just the same characteristic. But the peculiarity will be removed at once if the fourth column be placed immediately after the first : if this be done, *i.e.* if " red " be placed between " fair " and " brown " instead of at the end of the colour-series, the sign of the association in all the elementary tetrads will be the same. The colours will then run fair, red, brown, black, and this would seem to be the more natural order, considering the depth of the pigmentation.

Isotropic Contingency Tables.

5.16. A distribution of frequency of such a kind that the association in every elementary tetrad is of the same sign, possesses several useful and interesting properties, as shown in the following theorems. It will be termed an **isotropic** distribution.

(1) *In an isotropic distribution the sign of the association is the same not only for every elementary tetrad of adjacent frequencies, but for every set of four frequencies in the compartments common to two rows and two columns,* e.g. $(A_m B_n)$, $(A_{m+p}B_n)$, $(A_m B_{n+q})$, $(A_{m+p}B_{n+q})$.

For suppose that the sign of association in the elementary tetrads is positive, so that

$$(A_m B_n)(A_{m+1}B_{n+1}) > (A_{m+1}B_n)(A_m B_{n+1})$$

and similarly,

$$(A_{m+1}B_n)(A_{m+2}B_{n+1}) > (A_{m+2}B_n)(A_{m+1}B_{n+1})$$

Then multiplying up and cancelling, we have:

$$(A_m B_n)(A_{m+2}B_{n+1}) > (A_{m+2}B_n)(A_m B_{n+1})$$

That is to say, the association is still positive though the two columns A_m and A_{m+2} are no longer adjacent.

(2) *An isotropic distribution remains isotropic in whatever way it may be condensed by grouping together adjacent rows or columns.*

Thus from the first and third inequalities above we have, adding:

$$(A_mB_n)[(A_{m+1}B_{n+1}) + (A_{m+2}B_{n+1})] > (A_mB_{n+1})[(A_{m+1}B_n) + (A_{m+2}B_n)]$$

that is to say, the sign of the elementary association is unaffected by throwing the $(m+1)$th and $(m+2)$th columns into one.

(3) As the extreme case of the preceding theorem, we may suppose both rows and columns grouped and regrouped until only a 2×2-fold table is left; we then have the theorem:

If an isotropic distribution be reduced to a fourfold distribution in any way whatever, by addition of adjacent rows and columns, the sign of the association in such fourfold table is the same as in the elementary tetrads of the original table.

The case of complete independence is a special case of isotropy. For if

$$(A_mB_n) = (A_m)(B_n)/N$$

for all values of m and n, the association is evidently zero for every tetrad. Therefore the distribution remains independent in whatever way the table be grouped, or in whatever way the universe be limited by the omission of rows or columns. The expression " complete independence " is therefore justified.

From the work of the preceding section we may say that Table 5.2 is not isotropic as it stands, but may be regarded as a disarrangement of an isotropic distribution. It is best to rearrange such a table in isotropic order, as otherwise different reductions to fourfold form may lead to associations of different sign, though of course they need not necessarily do so.

5.17. The following will serve as an illustration of a table that is not isotropic, and cannot be rendered isotropic by any rearrangement of the order of rows and columns :—

TABLE 5.4. -*Showing the Frequencies of Different Combinations of Eye-colours in Father and Son.*

(Data of Sir F. Galton, from Karl Pearson, *Phil. Trans.*, A, vol. 195, 1900, p. 138; classification condensed.)

1. Blue. 2. Blue-green, grey. 3. Dark grey, hazel. 4. Brown.

Father's Eye-colour.

		1.	2.	3.	4.	Total.
Son's Eye-colour.	1	194	70	41	30	335
	2	83	124	41	36	284
	3	25	34	55	23	137
	4	56	36	43	109	244
	Total	358	264	180	198	1000

The following are the ratios of the frequency in column m to the sum of the frequencies in columns m and $m+1$:—

COLUMNS.

1 and 2.	2 and 3.	3 and 4.
0·735	0·631	0·577
0·401	0·752	0·532
0·424	0·382	0·705
0·609	0·456	0·283

The order in which the ratios run is different for each pair of columns, and it is accordingly impossible to make the table isotropic. The distribution of signs of association in the several tetrads is:

+	−	+
−	+	−
−	−	+

The distribution is a curious one, the associations in tetrads round the diagonal of the whole table being so markedly positive, and those in the immediately adjacent tetrads equally markedly negative. Neglecting the other signs, this is the effect that would be produced by taking an isotropic distribution and then increasing the frequencies in the diagonal compartments by a sufficient percentage. Comparison of the given table with others from the same source shows that the peculiarity is common to the great majority of the tables, and accordingly its origin demands explanation. Were such a table treated by the method of the contingency coefficient, or a similar summary method, alone, the peculiarity might not be remarked.

Complete Independence in Contingency Tables.

5.18. It may be noted that in the case of complete independence the distribution of frequency in every row is similar to the distribution in the row of totals, and the distribution in every column similar to that in the column of totals ; for in, say, the column A_n the frequencies are given by the relations:

$$(A_n B_1) = \frac{(A_n)}{N}(B_1), \qquad (A_n B_2) = \frac{(A_n)}{N}(B_2), \qquad (A_n B_3) = \frac{(A_n)}{N}(B)_3$$

and so on. This property is of special importance in the theory of variables.

Homogeneous and Heterogeneous Classification.

5.19. The classifications both of this and of the preceding chapters have one important characteristic in common, viz. that they are, so to speak, " homogeneous "—the principle of division being the same for all the sub-classes of any one class. Thus A's and a's are both subdivided into B's and β's, A_1's, A_2's, . . . A_s's into B_1's, B_2's, . . . B_t's, and so on. Clearly this is necessary in order to render possible those comparisons on which the discussions of associations and contingencies depend. If we only know that amongst the A's there is a certain percentage of B's, and amongst the a's a certain percentage of C's, there are no data for any conclusion.

Many classifications are, however, essentially of a heterogeneous character, *e.g.* biological classifications into orders, genera and species; the classifications of the causes of death in vital statistics and of occupations in the census. To take the last case as an illustration, the 1931 census of England and Wales divides occupations into 32 classes. Some of these are not further subdivided—*e.g.* " Fishermen." Others are subdivided into further general classes; *e.g.* Class 1 is divided into (1) Employers, (2) Furnacemen, (3) Foundry Workers, (4) Smiths, (5) Metal Machinists, (6) Fitters and (7) Other Workers. These sub-heads are necessarily peculiar to the class under which they occur and their number is arbitrary and variable, and different for each main heading ; but so long as the classification remains purely heterogeneous, however complex it may become, there is no opportunity for any discussion of causation within the limits of the matter so derived. *It is only when a homogeneous division is in some way introduced that we can begin to speak of associations and contingencies.*

5.20. This may be done in various ways according to the nature of the case. Thus the relative frequencies of different botanical families, genera or species may be discussed in connection with the topographical characters of their habitats—desert, marsh or heath—and we may observe statistical associations between given genera and situations of a given topographical type. The causes of death may be classified according to sex, or age, or occupation, and it then becomes possible to discuss the association of a given cause of death with one or other of the two sexes, with a given age-group or with a given occupation. Again, the classifications of deaths and of occupations are repeated at successive intervals of time ; and if they have remained strictly the same, it is also possible to discuss the association of a given occupation or a given cause of death with the earlier or later year of observation—*i.e.* to see whether the numbers of those engaged in the given occupation or succumbing to the given cause of death have increased or decreased. But in such circumstances the greatest care must be taken to see that the necessary condition as to the identity of the classifications at the two periods is fulfilled, and unfortunately it very seldom is fulfilled. All practical schemes of classification are subject to alteration and improvement from time to time, and these alterations, however desirable in themselves, render a certain number of comparisons impossible. Even where a classification has remained verbally the same, it is not necessarily really the same ; thus in the case of the causes of death, improved methods of diagnosis may transfer many deaths from one heading to another without any change in the incidence of the disease, and so bring about a virtual change in the classification. In any case, heterogeneous classification should be regarded only as a partial process, incomplete until a homogeneous division is introduced either directly or indirectly, *e.g.* by repetition.

Manifold Classification as a Series of Dichotomies.

5.21. From a theoretical point of view, manifold classification can be regarded as compounded of a series of dichotomies. Take, for example, a case we have already considered, that of the classification of a universe of men according to the eye-colours blue, grey, brown and green. We could have produced this fourfold division by three dichotomies. In fact,

dividing the universe first into those with blue eyes and those with not-blue eyes we get two classes. Then dividing again into those with brown eyes and those with not-brown eyes we get four classes. This operation on the class of blue-eyed men, however, results in one zero class, because there are no men with blue eyes which are at the same time brown, and one class which is, in fact, the class of blue-eyed men. Virtually, therefore, we have three classes: those with blue eyes, those with brown eyes, and the remainder. If we now dichotomise each of these into those with grey eyes and those with not-grey eyes, we shall again get, neglecting the zero classes, the four classes of the manifold classification.

5.22. It follows from this that any manifold classification can be regarded as produced by a succession of divisions in which, at each stage, each individual could fall into one of two alternatives, A or not-A.

Put in another way, this means that the possible answers to an unambiguous question can be reduced to a succession of answers of either " yes " or " no." For instance, suppose the question is, " How old are you, in years ? " We can replace this question by the succession of questions, " Are you one year old ? " " Are you two years old ? " . . . " Are you 120 years old ? " An answer of " 47 " to the first-mentioned question can then be expressed as an answer of " No " to the first 46 of these questions, " Yes " to the 47th and " No " to the rest.

Similarly, an answer to the question, " What is your name ? " can be reduced to the questions, " Is the first letter of your name A ? " " Is the first letter B ? " . . . " Is the second letter A ? " and so on. Replies to a more general question can be reduced to the same form by a convenient classification ; *e.g.* the replies to the question, " Are you in favour of war ? " can be classified in the four forms: " Favourable without qualification," " Favourable with some qualification," " Unfavourable without qualification," " Unfavourable with some qualification," and the answers to the questions can be reduced to answers " yes " or " no " to the questions, " Are you, without qualification, in favour of war ? " and so on.

Recording Classified Information on Punched Cards.

5.23. The information about an individual, considered as a member of a universe, is information whether he does or does not fall into the alternative classes which, as we have just seen, compose the most general homogeneous classification of the universe. If we imagine each individual filling in a questionnaire about himself, the totality of answers may, by suitably expressing the questions, be expressed as a number of " yes's " and " no's," and these replies express all the information about the individual.

This simple fact allows us to record the data in a most convenient way. Each individual is allotted a card, which is divided into a number of cells. Each cell corresponds to one of the dichotomies or simple questions the answers to which constitute the information. If the answer is " Yes," a hole is punched in the cell; if the answer is " No," the cell is left untouched.

The card of any individual will thus be like a complicated tram ticket, with holes punched in various places. The punching is usually performed either by hand with a ticket collector's punch, or with a machine similar in principle to the typewriter. The totality of punched cards forms a miniature of our universe—each individual has a card on which is recorded the whole of the information about him.

The use of this system lies in the fact that punched cards are easily handled and sorted by machinery. If, for example, we want to know a particular class-frequency, we can adjust certain electrical, pneumatic or mechanical stops, and the machine will segregate all the cards in the class and count them for us.

5.24. A similar device has been applied to the sorting of data by hand. A card is prepared with a row of circular holes punched all the way round near its edge, but so that no hole is open to the edge. Each hole corresponds to a dichotomy or a simple question. When preparing the card, if the individual falls into the A class, or the answer to the question is " Yes," a piece is clipped out of the card so that the hole is now open to the edge. If the individual falls into the not-A class, or the answer to the question is " No," the hole is left alone.

To separate the A's from the not-A's, or the " yes " cards from the " no " cards, they are arranged in a vertical plane so that corresponding cells are similarly placed. A skewer is then inserted in the appropriate hole and lifted. The not-A cards are lifted out, whilst the A cards fall away, since the piece of card between the hole and the edge has been cut away. By repeating the operation with the skewer in the appropriate holes we can isolate the cards in any given class. These can then be counted and the size of the class-frequency determined.

5.25. The labour of punching cards and the expense of machinery is justified only when the number of individuals is large and the number of ultimate classes is also large. This arises, for example, in the taking of a census of population.

Numerically Defined Attributes.

5.26. The attributes we have instanced in the foregoing pages have usually been of a qualitative kind. The methods described are, however, applicable to data classified on a numerical basis. Consider, for example, the following table :—

TABLE 5.5.—*Number of Families Deficient in Room Space in 95 Crowded London Wards.*

(Census of 1931, *Housing Report*, p. xxxii.)

Families deficient by	Standard Room Requirement (Rooms).							
	2	3	4	5	6	7	8	Total.
1 room	12,999	18,198	7,724	2,170	164	19	..	41,274
2 rooms	..	3,054	4,479	1,448	221	15	1	9,218
3 rooms	..	..	310	508	106	4	1	929
4 rooms	..	..	..	10	21	4	..	35
Total	12,999	21,252	12,513	4,136	512	42	2	51,456

The distinction between successive rows and columns is not quite of the kind of Table 5.2. In the latter, for instance, we drew a line between black

hair and brown, a line which could be drawn by anybody who was not colour-blind, although there may be border-line cases of mixed colours which would present difficulty. But in Table 5.5 above the line is drawn by counting—a much more precise operation. Moreover, the rows and columns have a certain natural order given by the numerical sequence. It would seem absurd to put the column which is headed " two rooms." between those headed " three rooms " and " four rooms," but in Table 5.2 there is no *a priori* reason for putting " black " between " brown " and " red."

5.27. We might also have a contingency table in which the attributes were measurable quantities, and the rows and columns of the table determined by ranges of those quantities. This, again, is slightly different from the case of the previous paragraph, for these ranges are to a large extent arbitrary, whereas in Table 5.5 the indivisible nature of the room compels us to count in units of at least one room.

5.28. Finally, we may have a table which is given by one qualitative attribute and one quantitative attribute. Consider, for example, the following :—

TABLE 5.6.—*Weight and Mentality in a Selection of Criminals.*

(Data from M. H. Whiting, "On the Association of Temperature, Pulse and Respiration with Physique and Intelligence in Criminals," *Biometrika*, vol. 11, pp. 1–37.)

Weight (lbs.).

		90–120.	120–130.	130–140.	140–150.	150 upward.	Totals.
Mentality.	Normal .	21	51	94	106	124	396
	Weak .	15	18	34	15	15	97
	Totals.	36	69	128	121	139	493

5.29. The methods of the previous chapters are applicable also to such tables. Numerically measurable quantities may, however, be treated by other methods, to which we shall come in due course. We mention the point here in order to remove any possible idea that the theory of attributes is concerned solely with qualitative classification, and is not appropriate to the more precise data given by a numerically assessable attribute.

SUMMARY.

1. The division of a universe according to an attribute A into a number of heads is called manifold classification. This is an extension of the idea of dichotomy, in which the universe is divided into two parts only.

2. Manifold classification according to two attributes A and B gives rise to a contingency table.

3. Association in a contingency table may be examined by reducing it in a number of ways to a 2×2 table.

4. The general nature of the association may be summarised by a coefficient.

5. We define

$$\delta_{mn} = (A_m B_n) - (A_m B_n)_0$$

The " square contingency " is given by:

$$\chi^2 = S\left\{\frac{\delta^2_{mn}}{(A_m B_n)_0}\right\} = S\left\{\frac{(A_m B_n)^2}{(A_m B_n)_0}\right\} - N$$

The " mean square contingency " by:

$$\phi^2 = \frac{\chi^2}{N}$$

6. Pearson's " coefficient of mean square contingency " is defined by:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} = \sqrt{\frac{\phi^2}{1 + \phi^2}}$$

7. Tschuprow's " coefficient of contingency " is defined by:

$$T^2 = \frac{\phi^2}{\sqrt{(s-1)(t-1)}}$$

8. Certain types of table, known as isotropic contingency tables, possess special features of some importance.

9. Any manifold classification may be regarded as a succession of dichotomies. This fact is the basis of the use of punched cards for recording and analysing statistical data.

10. Manifold classification may arise not only from an attribute which is specified under heads of a qualitative kind, but also from a quantitative attribute specified by counting or measurement.

EXERCISES.

5.1. (Data from Karl Pearson, "On the Inheritance of the Mental and Moral Characters in Man," *Jour. of the Anthrop. Inst.*, vol. 33, and *Biometrika*, vol. 3.) Find the coefficient of contingency (coefficient of mean square contingency) for the two tables below, showing the resemblance between brothers for athletic capacity and between sisters for temper. Show that neither table is even remotely isotropic. (As stated in 5.11, the coefficient of contingency should not as a rule be used for tables smaller than 5×5-fold: these small tables are given to illustrate the method, while avoiding lengthy arithmetic.)

A. ATHLETIC CAPACITY.

First Brother.

				Athletic.	Betwixt.	Non-athletic.	Total.
Second Brother	Athletic	.	.	0ŦI	20	140	1066
	Betwixt	.	.	0ᴢ	76	9	105
	Non-athletic	.	.	906	9	370	519
	Total	.		1066	105	519	1690

B. TEMPER.

First Sister.

	Quick.	Good-natured.	Sullen.	Total.
Quick	198	177	77	452
Good-natured . .	177	996	165	1338
Sullen . . .	77	165	120	362
Total .	452	1338	362	2152

(Second Sister, as row label on left side)

5.2. Calculate T and C for the following table, and trace the association between the progress of building and the urban character of the district :—

Houses in England and Wales. (*Census of 1901. Summary Table X.*)
(*000's omitted.*)

	Inhabited.	Uninhabited.	Building.	Total.
Adm. County of London .	571	40	5	616
Other urban districts . .	4064	285	45	4394
Rural districts . . .	1625	124	12	1761
Total for England and Wales	6260	449	62	6771

5.3. Show that for a given s and t, C and T are equal for two values of ϕ^2, one of which is zero; that for ϕ^2 between these values $C > T$; and that for ϕ^2 greater than the higher value $T > C$.

5.4. Find whether the following contingency table is isotropic, and if it is not, ascertain whether it can be arranged in an isotropic form:—

	$A_1.$	$A_2.$	$A_3.$	$A_4.$	$A_5.$	Totals.
B_1	90	43	17	27	16	193
B_2	235	88	44	60	40	467
B_3	300	103	54	71	48	576
Totals	625	234	115	158	104	1236

5.5. Calculate C and T for the table of the previous example.

5.6. Show that in a positively isotropic contingency table,

$$\frac{\delta_{11}}{(A_1B_1)_0} > \frac{\delta_{1s}}{(A_1B_s)_0} \quad \text{and is} \quad > \frac{\delta_{r1}}{(A_rB_1)_0}$$

5.7. 1000 subjects of English, French, German, Italian and Spanish nationality were asked to name their preference among the music of those five

nationalities. The results were as follows (1 = English, 2 = French, 3 = German, 4 = Italian, 5 = Spanish) :—

Nationality of Music Preferred.

		1.	2.	3.	4.	5.	Totals.
Nationality of Subject.	1	32	16	75	47	30	200
	2	10	67	42	41	40	200
	3	12	23	107	36	22	200
	4	16	20	44	76	44	200
	5	8	53	30	43	66	200
	Totals	78	179	298	243	202	1000

Discuss the association between the nationality of the subject and the nationality of the music preferred.

5.8. In Table 5.6 calculate C and T, and discuss the light thrown by this table on the association between physique and intelligence in the criminals of the data.

5.9. Show that for a 2×2 contingency table in which the frequencies are $(A_1 B_1) = a$, $(A_2 B_1) = b$, $(A_1 B_2) = c$ and $(A_2 B_2) = d$,

$$\chi^2 = \frac{(a + b + c + d)(ad - bc)^2}{(a + b)(c + d)(b + d)(a + c)}$$

and hence find C and T in terms of a, b, c, d.

5.10. In a paper discussing whether laterality of hand is associated with laterality of eye (measured by astigmatism, acuity of vision, etc.) T. L. Woo obtained the following results (*Biometrika*, vol. 20A, pp. 79–148): -

Ocular Laterality for General Astigmatism.

		"Left-eyed."	Ambiocular.	"Right-eyed."	Totals.
Manual Laterality as Determined by a Balancing Test.	Left-handed .	34	62	28	124
	Ambidextrous	27	28	20	75
	Right-handed	57	105	52	214
	Totals .	118	195	100	413

Show that laterality of eye is only slightly associated with laterality of hand.

CHAPTER 6.

FREQUENCY-DISTRIBUTIONS.

Variables.

6.1. As we emphasised at the close of the last chapter, the methods of the theory of attributes are applicable to all observations, whether qualitative or quantitative. We have now to proceed to the consideration of special processes adapted to the treatment of quantitative data, but not as a rule available for the discussion of purely qualitative observations (though there are some important exceptions to this statement, as suggested in **1.2**).

Numerical measurement is applied only to a quantity which can present more than one numerical value. Otherwise there would be no point in measuring it. Such a quantity is therefore called a **variable**,[1] and this section of our work may be termed the **theory of variables.**

As common examples of variables which are subject to statistical treatment we may cite birth- and death-rates, prices, wages, barometer readings, rainfall records, and measurements or enumerations (*e.g.* of glands, spines or petals) on animals or plants.

Quantities which can take any numerical value within a certain range are called **continuous variables.** Such, for example, are birth-rates and barometric readings. Quantities which can take only discrete values are called **discontinuous variables.** This class, for instance, would include data of the number of petals on flowers or the number of rooms in a house.

Frequency-distributions.

6.2. If some hundreds or thousands of values of a variable have been noted merely in the arbitrary order in which they occur, the mind cannot properly grasp the significance of the record. We must condense the data by some method of ranking or classification before their characteristics can be comprehended.

One way of doing this would be to dichotomise the data by classifying the individuals as A's or not-A's, according as the value of the variable exceeded or fell short of some given value. But this is too crude, and the sacrifice of information is too great. A manifold classification, however, avoids the crudity of the dichotomous form, since the classes may be made as numerous as we please. Moreover, numerical measurements lend themselves with peculiar readiness to a manifold classification, for the class limits can be conveniently and precisely defined by assigned values of the variable. -

6.3. For convenience, the values of the variable chosen to define the successive classes should be equidistant, so that the numbers of observations in different classes are comparable.

---

[1] It is also called a *variate.* We shall use the two terms as synonymous.

The interval chosen for classifying is called the **class-interval**, and the frequency in a particular class-interval is called a **class-frequency**.

Thus, for measurements of stature, the class-interval might be 1 inch, or 2 centimetres, and the class-frequencies would be the numbers of individuals whose statures fell within each successive inch or each successive 2 centimetres of the scale; returns of birth- or death-rates might be grouped to the nearest unit per thousand of the population ; returns of wages might be classified to the nearest shilling, or, if it is desired to obtain a more condensed table, to the nearest five or ten shillings. Discontinuous variables to a great extent determine their own class-intervals, which must either be equal in width to the unit amount of variation, or equal to some multiple of it. For example, in enumerations of the number of rooms in a house we naturally take our class-interval to be one room ; in enumerations of the petals on a flower we may take one petal, or, if the range of variation is very great, say five petals or more.

6.4. The manner in which the class-frequencies are distributed over the class-intervals is spoken of as the **frequency-distribution** of the variable.

A few illustrations will make clearer the nature of such frequency-distributions, and the service which they render in summarising a long and complex record.

(a) Table 6.1. In this illustration the birth-rates per thousand of the population in 1933 of 1567 local government areas of England have been classified to the nearest unit; i.e. the number of districts has been counted in which the birth-rate was between 1·5 per thousand and 2·5, between 2·5 and 3·5, and so on. The frequency-distribution is shown by the table.

TABLE 6.1.—*Showing the Number of Local Government Areas in England with Specified Birth-rates per Thousand of Population.* (Material from the Registrar-General's Statistical Review of England and Wales for 1933.)

Birth-rate.	Number of Districts with Birth-rate Between Limits Stated.	Birth-rate.	Number of Districts with Birth-rate Between Limits Stated.
1·5– 2·5	1	13·5–14·5	271
2·5– 3·5	2	14·5–15·5	190
3·5– 4·5	2	15·5–16·5	127
4·5– 5·5	3	16·5–17·5	89
5·5– 6·5	7	17·5–18·5	78
6·5– 7·5	9	18·5–19·5	37
7·5– 8·5	14	19·5–20·5	21
8·5– 9·5	41	20·5–21·5	17
9·5–10·5	83	21·5–22·5	4
10·5–11·5	131	22·5–23·5	4
11·5–12·5	192	23·5–24·5	2
12·5–13·5	242	Total	1567

Although a glance through the original returns, which are spread amongst many other figures over 42 pages, fails to convey any definite impression,

a brief inspection of the above table brings out a number of important points. Thus, we see that the birth-rates range, in round numbers, from 2 to 24 per thousand ; that the birth-rates in some 75 per cent. of the districts lie within the narrow limits 10·5 to 16·5, the rates most frequent being near 14 ; and so on. It may be remarked that some of the areas are very small, with no more than 10 or 20 births, and these account mainly for the extremely divergent rates.

(b) Table 6.2. The numbers of stigmatic rays on a number of Shirley poppies were counted. As the range of variation is not great, the unit is taken as the class-interval. The frequency-distribution is given by the following table :—

TABLE 6.2.—*Showing the Frequencies of Seed Capsules on certain Shirley Poppies, with Different Numbers of Stigmatic Rays.* (Cited from *Biometrika*, vol. 2, 1902, p. 89.)

Number of Stigmatic Rays.	Number of Capsules with said Number of Stigmatic Rays.	Number of Stigmatic Rays.	Number of Capsules with said Number of Stigmatic Rays.
6	3	14	302
7	11	15	234
8	38	16	128
9	106	17	50
10	152	18	19
11	238	19	3
12	305	20	1
13	315		
		Total	1905

The numbers of rays range from 6 to 20, the most usual numbers being 12, 13 or 14.

(c) Table 6.3. 206 screws were taken as they came off the lathe which was turning them. Their lengths, which should have been 1 inch, were measured. The following table shows the screws classified by the number of thousandths of an inch by which they exceeded or fell short of 1 inch in length : —

TABLE 6.3.—*Showing the Frequencies of Screws Classified according to the Extent to which they Varied in Length from the Standard of 1 Inch.* (Unpublished data, A. M. Lester.)

Difference in Length from 1 Inch (Thousandths of an Inch).	Number of Screws.	Difference in Length from 1 Inch (Thousandths of an Inch).	Number of Screws.
−6 to −5	1	+1 to +2	34
−5 to −4	4	+2 to +3	25
−4 to −3	11	+3 to +4	16
−3 to −2	22	+4 to +5	8
−2 to −1	25	+5 to +6	1
−1 to 0	27		
0 to +1	32	Total	206

It will be seen that the maximum frequency, *i.e.* 34, occurs for screws from 0·001 to 0·002 inch in excess of the standard. About 80 per cent. lie in the range three-thousandths of an inch on either side of the standard.

6.5. Expanding slightly the brief description we have given, tables setting out frequency-distributions are formed in the following way :—

.(1) The magnitude of the class-interval is first fixed. In Tables 6.1, 6.2 and 6.3 one unit was chosen.

(2) The position or origin of the intervals must then be determined ; *e.g.* in Table 6.1 we must decide whether to take as intervals 9–10, 10–11, 11–12, etc., or 9·5–10·5, 10·5–11·5, 11·5–12·5, etc.

(3) This choice having been made, the complete scale of intervals is fixed and the observations are classified accordingly.

(4) The process of classification being finished, a table is drawn up on the general lines of Tables 6.1–6.3, showing the total number of observations in each class-interval.

It is necessary to make a few remarks about each of these heads.

Magnitude of Class-interval.

6.6. As already remarked, in cases where the variation proceeds by discrete steps of considerable magnitude as compared with the range of variation, there is very little choice as regards the magnitude of the class-interval. The unit will in general have to serve. But if the variation be continuous, or at least take place by discrete steps which are small in comparison with the whole range of variation, there is no such natural class-interval, and its choice is a matter for judgment.

The two conditions which guide the choice are these : (*a*) We desire to be able to treat all the values assigned to any one class, without serious error, as if they were equal to the mid-value of the class-interval, *e.g.* as if the birth-rate of every district in the first class of Table 6.1 were exactly 2·0, the birth-rate of every district in the second class 3·0, and so on ; (*b*) for convenience and brevity we desire to make the interval as large as possible, subject to the first condition. These conditions will generally be fulfilled if the interval be so chosen that the whole number of classes lies between 15 and 25. A number of classes less than, say, ten leads in general to very appreciable inaccuracy, and a number over, say, thirty makes a somewhat unwieldy table. A preliminary inspection of the record should accordingly be made and the highest and lowest values be picked out. Dividing the difference between these by, say, twenty-five, we have an approximate value for the interval. The actual value should be the nearest integer or simple fraction.

Position of·Intervals.

6.7. The position or starting-point of the intervals is, as a rule, more or less a matter of indifference. It can therefore be chosen as is most convenient for the particular case under discussion, *e.g.* so that the limits of the intervals are integers, or, as in Table 6.1, so that the mid-values are integers. It may also be chosen so that no limits correspond exactly to any recorded value of the variate, in order to obviate any difficulty in deciding to which class a particular individual should be assigned (*cf.* **6.9**).

The location of the intervals is, however, important when the values of the variate tend for some reason to cluster round particular values. Such a case arises, for instance, in age returns, owing to the tendency to state a round number where the true age is unknown, or a reluctance to admit one's real age.[1] It is also common wherever there is some doubt as to the final digit in reading a scale, and scope is given to the idiosyncrasies of the observer.

Table 6.4 shows results for four observers as illustrations, the frequencies being reduced for comparability to a total of 1000. Column A is based on measures by G. U. Yule, on drawings, to the nearest tenth of a millimetre. It is recognised, of course, that measures cannot really

TABLE 6.4.— *Frequency-distributions of Final Digits in Measurements by Four Observers.* (G. U. Yule, "On Reading a Scale," *Journal Royal Statistical Society*, vol. 90, 1927, p. 570.)

Final Digit.	Frequency of Final Digit per 1000.			
	A.	B.	C.	D.
0	158	122	251	358
1	97	98	37	49
2	125	95	80	90
3	73	90	72	63
4	76	100	55	37
5	71	112	222	211
6	90	98	71	62
7	56	99	75	70
8	126	101	72	44
9	129	81	65	16
Total	1001	999	1000	1000
Actual observations }	1258	3000	1000	1000

be made to such a degree of precision ; but the measurer believed that he was making them carefully, and as they were made with a Zeiss scale, in which the divisions are ruled on the under side of a piece of plate-glass, readings were unaffected by parallax. Nevertheless, it will be seen that the zeros, and also 2, 8 and 9, were heavily over-emphasised—an odd selection of preferences! On the whole, the centre of the millimetre was neglected and measures piled up at the two ends.

The data for columns B, C and D are all drawn from the same published report, and refer to sundry head measurements taken on the living subject. On the basis of a statement in the introduction to the report, it was possible to compile the data separately for the three assistants (B, C, D) who had done the actual measuring. It will be seen that B was rather good : there is a relatively slight excess at 0 and 5, but otherwise his measurements are

[1] This effect is practically the same for men as for women. *Cf.* Table I in the Appendix to the paper cited in the heading to Table 6.4 above.

fairly uniformly distributed. C was decidedly not good, rounding off nearly one measurement in two to the nearest centimetre or half-centimetre. D was simply outrageously bad—so bad that it might have been better not to publish his measurements. Nearly 57 per cent. of his measurements were made only to the nearest centimetre or half-centimetre—a quite inadequate degree of precision for head measurements often only a few centimetres in magnitude.

When there is any possibility of clustering of variate values, it is as well to subject the data to a close examination before finally fixing on the method of classification. On the whole, the intervals should be arranged as far as possible so that the values round which the clustering occurs fall towards the interval mid-values. This procedure avoids sensible error in the assumption that the interval mid-value is approximately representative of the values of the class.

Classification.

6.8. The scale of intervals having been fixed, the observations may be classified. If the number of observations is not large, it will be sufficient to mark the limits of successive intervals in a column down the left-hand side of a sheet of paper, and transfer the entries of the original record to this sheet by marking a 1 on the line corresponding to any class for each entry assigned thereto. It saves time in subsequent totalling if each fifth entry in a class is marked by a diagonal across the preceding four, or by leaving a space.

The disadvantage in this process is that it offers no facilities for checking : if a repetition of the classification leads to a different result, there is no means of tracing the error. If the number of observations is at all considerable and accuracy is essential, it is accordingly better to enter the values observed on cards, one to each observation. These are then dealt out into packs according to their classes, and the whole work checked by running through the pack corresponding to each class, and verifying that no cards have been wrongly sorted.

6.9. In some cases difficulties may arise in classifying, owing to the occurrence of observed values corresponding to class-limits. Thus, in compiling Table 6.1 some districts will have been noted with birth-rates entered in the Registrar-General's returns as 16·5, 17·5 or 18·5, any one of which might at first sight have been apparently assigned indifferently to either of two adjacent classes. In such a case, however, where the original figures for numbers of births and population are available, the difficulty may be readily surmounted by working out the rate to another place of decimals : if the rate stated to be 16·5 proves to be 16·502, it will be sorted to the class 16·5–17·5 ; if 16·498, to the class 15·5–16·5. Birth-rates that work out to half-units exactly do not occur in this example, and so there is no real difficulty.

In the case of Table 6.3, again, there is little difficulty in knowing the class to which an individual should be assigned.

Difficulties of this type may, in fact, always be avoided if they are borne in mind in fixing the class-intervals, by fixing the intervals to a further place of decimals or a smaller fraction than the values in the original record. Thus, if statures are measured to the nearest centimetre, the class-intervals may be taken as 150·5–151·5, 151·5–152·5, etc. ; if to the

nearest eighth of an inch, the intervals may be $59\frac{15}{16}$–$60\frac{1}{16}$, $60\frac{1}{16}$–$61\frac{1}{16}$, and so on.

If the difficulty is not evaded in any of these ways, it is usual to assign one-half of an intermediate observation to each adjacent class, with the result that half-units occur in the class-frequencies (*cf.* Table 6.9, p. 98). The procedure is rough, but probably good enough for practical purposes ; strict precision is usually unattainable, for in point of fact the odd way in which different individuals read a scale, for example, renders it impossible to assign exact limits to intervals.

Tabulation.

6.10. As regards the actual drafting of the final table there is little to be said, except that care should be taken to express the class-limits clearly and, if necessary, to say how the difficulty of intermediate values has been met or evaded. The class-limits are perhaps best given as in Tables 6.1 and 6.3, but may be more briefly indicated by the mid-values of the class-intervals. Thus, Table 6.1 might have been given in the form :

Birth-rate per 1000 to the Nearest Unit.	Number of Districts with said Birth-rate.
2	1
3	2
4	2
etc.	etc.

It should be noticed that the method of defining class-intervals adopted in Table 6.3 leaves the class-limits uncertain unless the degree of accuracy of the measurements is also given. Thus, in a table giving frequencies of men in certain height-ranges of 1 inch in width, say " 57 and less than 58," etc., if measurements were taken to the nearest eighth of an inch, the class-limits are really $56\frac{15}{16}$–$57\frac{15}{16}$, $57\frac{15}{16}$–$58\frac{15}{16}$, etc. ; if they were only taken to the nearest quarter of an inch, the limits are $56\frac{7}{8}$–$57\frac{7}{8}$, $57\frac{7}{8}$–$58\frac{7}{8}$, etc. With such a form of tabulation a statement as to the number of significant figures in the original record is therefore essential. It is better, perhaps, to state the true class-limits and avoid ambiguity.

6.11. The rule that class-intervals should be all equal is one that is very frequently broken in official statistical publications, principally in order to condense an otherwise unwieldy table, thus not only saving space in printing but also considerable expense in compilation, or possibly, in the case of confidential figures, to avoid giving a class which would contain only one or two observations, the identity of which might be guessed. It would hardly be legitimate, for example, to give a return of incomes relating to a limited district in such a form that the income of the two or three wealthiest men in the district would be clear to any intelligent reader with local knowledge.

If the class-intervals be made unequal, the application of many statistical methods is rendered awkward, or even impossible. Further, the relative values of the frequencies are misleading, so that the table is not perspicuous. Thus, consider the first two columns of Table 6.5, showing

the number of persons liable to sur-tax and super-tax classified according to their annual income. On running the eye down the column headed " Number of Persons," the attention is at once caught by the three irregularities at the classes " £3000 and not exceeding £4000," " £8000 and not exceeding £10,000," and " £10,000 and not exceeding £15,000." But these have no real significance ; they are merely due to changes in the magnitude of the class-interval at those points. A further change occurs at the £30,000 and at the £50,000 mark, although the attention is not directed thereto by any marked irregularity in the frequencies.

TABLE 6.5.—*Showing the Numbers of Persons in the United Kingdom liable to Sur-tax and Super-tax in the Year beginning 5th April 1931, classified according to the Magnitude of their Annual Income.* (From the Statistical Abstract for the United Kingdom for the Years 1913 and 1919–32, Cmd. 4489.)

Annual Income (£000).			Number of Persons.	Frequency per £500 Interval.
2 and not exceeding		2·5	23,988	23,988
2·5	,,	3	15,781	15,781
3	,,	4	17,979	8,989
4	,,	5	9,755	4,877
5	,,	6	5,921	2,960
6	,,	7	3,729	1,864
7	,,	8	2,546	1,273
8	,,	10	3,193	798
10	,,	15	3,616	362
15	,,	20	1,328	133
20	,,	25	679	68
25	,,	30	378	38
30	,,	40	372	19
40	,,	50	192	10
50	,,	75	182	4
75	,,	100	57	1
100 and over			94	?
Total number of persons			89,790	—

To make the class-frequencies really comparable *inter se* they must first be reduced to a common interval as basis, say £500, by dividing the third and subsequent numbers by 2, the eighth by 4, and so on. This gives the mean frequencies tabulated in the third column of Table 6.5. The reduction is, however, impossible in the case of the last class, for we are told only the number of persons with an income of £100,000 and upwards. Such an indefinite class is in many respects a great inconvenience, and should always be avoided in work not subjected to the necessary limitations of official publications.

6.12. The general rule that intervals should be equal must not be held to bar the analysis by smaller equal intervals of some portion of the range over which the frequency varies very rapidly. In Table 6.11, page 100, for example, giving the numbers of deaths from scarlet fever at successive ages, it is desirable to give the numbers of deaths in each year for the first five years, so as to bring out the rapid rise to the maximum in the third year of life.

Graphical Representation: Frequency-polygon and Histogram.

6.13. It is often convenient to represent the frequency-distribution by means of a diagram which conveys to the eye the general run of the observations. The following short table, giving the distribution of head-breadths for 1000 men, will serve as an example :—

TABLE 6.6.—*Showing the Frequency-distribution of Head-breadths for Students at Cambridge. Measurements taken to the nearest Tenth of an Inch.* (Cited from W. R. Macdonell, *Biometrika*, vol. 1, 1902, p. 220.)

Head-breadth in Inches.	Number of Men with said Head-breadth.	Head-breadth in Inches.	Number of Men with said Head-breadth.
5·5	3	6·3	99
5·6	12	6·4	37
5·7	43	6·5	15
5·8	80	6·6	12
5·9	131	6·7	3
6·0	236	6·8	2
6·1	185		
6·2	142	Total	1000

Taking a piece of squared paper ruled, say, in inches and tenths, mark off along a horizontal base-line a scale representing class-intervals ; a half-inch to the class-interval would be suitable. Then choose a vertical scale for the class-frequencies, say 50 observations per interval to the inch, and mark off, on the verticals or *ordinates* through the points marked 5·5, 5·6, 5·7, . . . at the centres of the class-intervals on the base-line, heights representing on this scale the class-frequencies 3, 12, 43, . . . The diagram may then be completed in one of two ways : (1) as a **frequency-polygon**, by joining up the marks on the verticals by straight lines, the last points at each end being joined down to the base at the centre of the next class-interval (fig. 6.1) ; or (2) as a column diagram or **histogram**, short horizontals being drawn through the marks on the verticals (fig. 6·2), which now form the central axes of a series of rectangles representing the class-frequencies.

6.14. The student should note that in any such diagram, of either form, a certain *area* represents a given number of observations. On the scales suggested, 1 inch on the horizontal represents 2 intervals, and 1 inch on the vertical represents 50 observations per interval : 1 square inch therefore represents $50 \times 2 = 100$ observations. The diagrams are, however, conventional : in both cases the whole area of the figure is proportional to the total number of observations, but the area over every interval is not correct in the case of the frequency-polygon, and the frequency of every fraction of any interval is not the same, as suggested by the histogram. The area shown by the frequency-polygon over any interval with an ordinate y_2 (fig. 6.3) is only correct if the tops of the three successive ordinates y_1, y_2, y_3 lie on a line, *i.e.* if $y_2 = \frac{1}{2}(y_1 + y_3)$, the areas of the two little triangles shaded in the figure being equal. If y_2 fall short of this value, the area shown by the polygon is too great ; if y_2 exceed it, the area shown by the polygon is too small ; and if, for this reason, the

frequency-polygon tends to become very misleading at any part of the range, it is better to use the histogram.
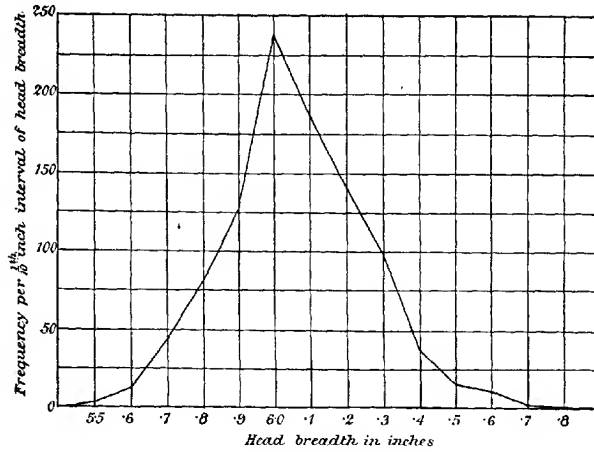
FIG. 6.1.– Frequency-polygon for Head-breadths of 1000 Cambridge Students. (Table 6.6.)
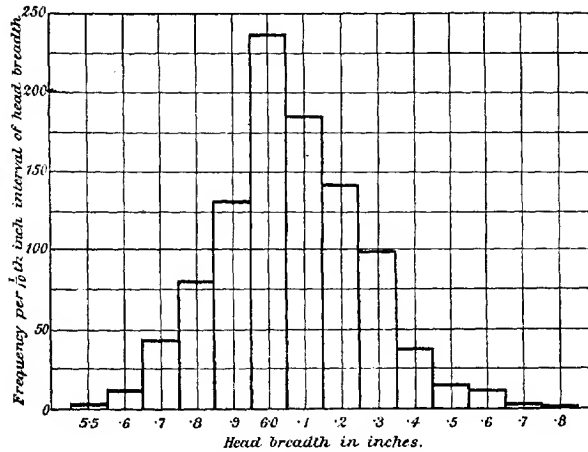
FIG. 6.2.—Histogram for the same data as fig. 6.1.

6.15. The histogram may also be used when the class-intervals are unequal. The construction of the previous section is easily adapted to

such cases. All that is necessary is to describe an area equal, on the scale adopted, to the frequency in a particular interval ; this is done, as before, by erecting at the centre of the interval an ordinate equal in length to

FIG. 6.3.

the total frequency divided by the width of the interval.
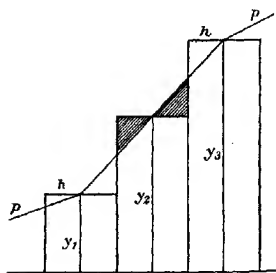
An example of this kind of construction is given in fig. 6.11 (Table 6.11). The frequencies of deaths for ages over 5 years are given in 5-yearly periods, whereas those for ages under 5 years are given in 1-yearly periods. On the scale indicated, therefore, the height of the cell of the histogram corresponding to the ages 2–3 years is 89, the class-frequency ; that of the cell corresponding to the ages 5–10 is 42·6, *i.e.* 213 divided by 5. Hence the areas of the two cells are, to the scale adopted, 89 and 213, respectively, so that the areas accurately represent the frequencies.

Frequency-curves.

6.16. If the class-intervals be made smaller, and at the same time the number of observations increased so that the class-frequencies may remain finite, the polygon and the histogram will approach more and
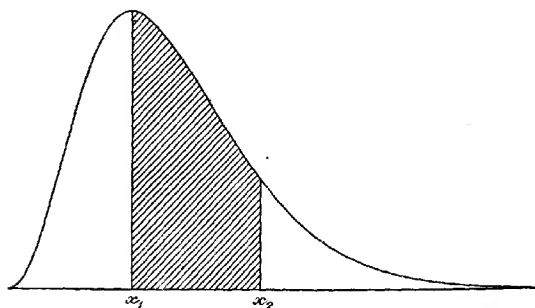
FIG. 6.4.

more closely to a smooth curve. Such an ideal limit to the polygon or the histogram is called a **frequency-curve**. It is a concept of supreme importance in statistical theory.

In the frequency-curve the area between any two ordinates whatever is proportional to the number of observations falling between the corresponding values of the variable. Thus, the number of observations falling between the values of the variable x_1 and x_2 in fig. 6.4 will be proportional to the area of the shaded strip in the figure ; the number of

observed values greater than x_2 will be given by the area of the curve to the right of the ordinate at x_2; and so on.

6.17. When we come to consider the theory of sampling we shall regard the frequency curve as representing a universe from which the actual data are a specimen. The frequency-polygon and the histogram will then be approximations to the curve, but will diverge from it to some extent owing to fluctuations of sampling. For the present we must defer a closer inquiry into this subject. We may remark, however, that when the number of observations is considerable—say a thousand at least—the run of the class-frequencies is usually sufficiently smooth to give a good notion of the form of the " ideal " distribution.

Some Common Types of Frequency-distribution.

6.18. The forms presented by smoothly running sets of data are almost endless in their variety, but among them we may notice a comparatively small number of simple types. Such types also form a set into which more complex distributions may often be analysed. For elementary purposes it is sufficient to consider four fundamental simple types, which we shall call the symmetrical distribution, the moderately asymmetrical or skew distribution,[1] the extremely asymmetrical or J-shaped distribution and the U-shaped distribution. In the following sections we give some examples of each of these types, together with a few more complex distributions.

The Symmetrical Distribution.

6.19. In this type the class-frequencies decrease to zero symmetrically on either side of a central maximum. Fig. 6.5 illustrates the ideal form of the distribution.
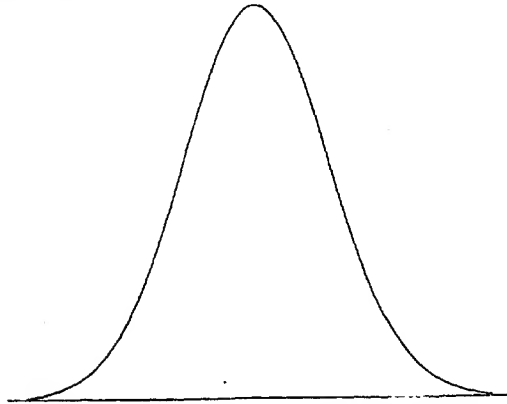
Fig. 6.5.—An Ideal Symmetrical Frequency-distribution.

---

[1] These two types, from their shape, are frequently referred to as "humped," "cocked hat," "single peaked," and so on.

Being a special case of the more general type described under the second heading, this form of distribution is comparatively rare. It occurs in the case of biometric, more especially anthropometric, measurements, from which the following illustration is drawn, and is important in much theoretical work. Table 6.7 shows the frequency-distribution of statures for adult males born in the British Isles, from data published by a

TABLE 6.7.—*Showing the Frequency-distributions of Statures for Adult Males born in England, Scotland, Wales and Ireland.* (Final Report of the Anthropometric Committee to the British Association.) (*Report*, 1883, p. 256.) *As Measurements are stated to have been taken to the nearest $\frac{1}{8}$th of an Inch, the Class-intervals are here presumably* $56\frac{15}{16}$–$57\frac{15}{16}$, $57\frac{15}{16}$–$58\frac{15}{16}$, *and so on (cf. 6.9).* (See fig. 6.6.)

Height without shoes, Inches.	Number of Men within said Limits of Height. Place of Birth—				Total.
	England.	Scotland.	Wales.	Ireland.	
57–	1	—	1	—	2
58–	3	1	—	—	4
59–	12	—	1	1	14
60–	39	2	—	—	41
61–	70	2	9	2	83
62–	128	9	30	2	169
63–	320	19	48	7	394
64–	524	47	83	15	669
65–	740	109	108	33	990
66–	881	139	145	58	1223
67–	918	210	128	73	1329
68–	886	210	72	62	1230
69–	753	218	52	40	1063
70–	473	115	33	25	646
71–	254	102	21	15	392
72–	117	69	6	10	202
73–	48	26	2	3	79
74–	18	15	1	—	32
75–	9	6	1	—	16
76–	1	4	—	—	5
77–	1	1	—	—	2
Total	6194	1304	741	346	8585

British Association Committee in 1883, the figures being given separately for persons born in England, Scotland, Wales and Ireland, and totalled in the last column. These frequency-distributions are approximately of the symmetrical type. The frequency-polygon for the totals given by the last column of the table is shown in fig. 6.6. The student will notice that an error of $\frac{1}{16}$ inch, scarcely appreciable in the diagram on its reduced scale, is neglected in the scale shown on the base-line, the intervals being treated as if they were 57–58, 58–59, etc. Diagrams should be drawn for comparison showing, to a good open scale, the separate distributions for England, Scotland, Wales and Ireland.

The Moderately Asymmetrical (Skew) Distribution.

6.20. In this case the class-frequencies decrease with markedly greater rapidity on one side of the maximum than on the other, as in

fig. 6.7 (*a*) or (*b*). This is the most common of all smooth forms of frequency-distribution, illustrations occurring in statistics from almost
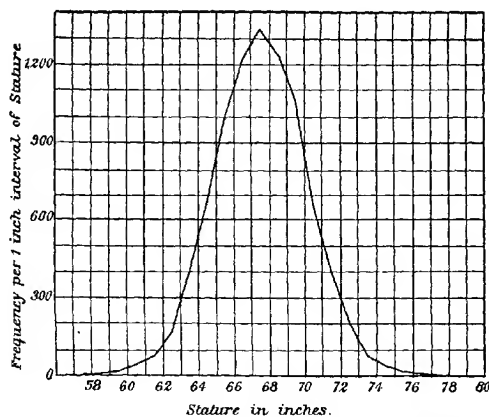
FIG. 6.6.—Frequency-distribution of Stature for 8585 Adult Males born in the British Isles. (Table 6.7.)
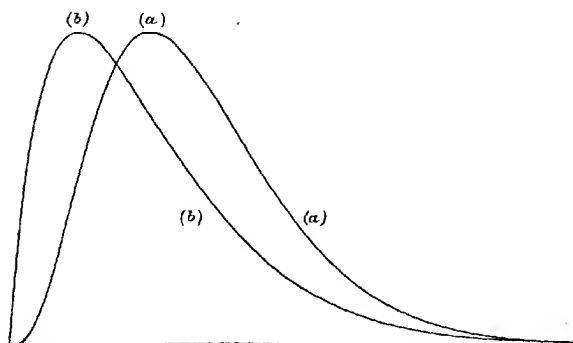
FIG. 6.7.—Ideal Distributions of the Moderately Asymmetrical Form.

every source. The distribution of birth-rates given in Table 6.1 is slightly asymmetrical.

The distribution of Australian marriages given in Table 6.8 (fig. 6.8) is rather more asymmetrical and is of the type (*a*) of fig. 6.7. The frequency attains its maximum for ages between 24 and 27 and then tails off slowly. We have not drawn the tail of the curve, which is very close to the x-axis, for values of the variate above 58·5.

TABLE 6.8.—*Showing Numbers of Marriages Contracted in Australia, 1907–14, arranged according to the Age of Bridegroom in 3-Year Groups.* (From S. J. Pretorius, "Skew Bivariate Frequency Surfaces," *Biometrika*, vol. 22, 1930–31, p. 210.) (See fig. 6.8.)

Age of Bridegroom (Central Value of 3-Year Range, in Years).	Number of Marriages.	Age of Bridegroom (Central Value of 3-Year Range, in Years).	Number of Marriages.
16·5	294	55·5	1,655
19·5	10,995	58·5	1,100
22·5	61,001	61·5	810
25·5	73,054	64·5	649
28·5	56,501	67·5	487
31·5	33,478	70·5	326
34·5	20,569	73·5	211
37·5	14,281	76·5	119
40·5	9,320	79·5	73
43·5	6,236	82·5	27
46·5	4,770	85·5	14
49·5	3,620	88·5	5
52·5	2,190		
		Total	301,785

Table 6.9 and fig. 6.9 give a biological illustration, viz. the distribution of fecundity (ratio of yearling foals produced to coverings) in mares. The student should notice the difficulty of classification in this case : the class-interval chosen throughout the middle of the range is 1/15th, but the last interval is " 29/30–1." This is not a whole interval, but it is more than a half, for all the cases of complete fecundity are reckoned into the class. In the diagram (fig. 6.9) it has been reckoned as a whole class, and this gives a smooth distribution.

· To take an illustration from meteorology, the distribution of barometer heights at any one station over a period of time is, in general, asymmetrical, the most frequent heights lying towards the upper end of the range for stations in England and Wales. Table 6.10 and fig. 6.10 show the distribution for daily observations at Greenwich during the years 1848–1926 inclusive.

The distributions of Tables 6.8–6.10 all follow more or less the type of fig. 6.7 (*a*), the frequency tailing off, at the steeper end of the distribution, in such a way as to suggest that the ideal curve is tangential to the base. Cases of greater asymmetry, suggesting an ideal curve that meets the base (at one end) at a finite angle, even a right angle, as in fig. 6.7 (*b*), are less frequent, but occur occasionally. The distribution of deaths from scarlet fever, according to age, affords one such example of a more asymmetrical kind. The actual figures for this case are given in Table 6.11 and illustrated by fig. 6.11 ; and it will be seen that the frequency of deaths reaches a maximum for children aged " 2 and under 3," the number rising very rapidly to the maximum, and thence falling so slowly that there is still an appreciable frequency for persons over 50 years of age.

Asymmetrical curves are also said to be " skew." In Chapter 9

we shall consider skewness at some length and discuss various ways of measuring it.   In particular we shall find that skewness has a sign, and
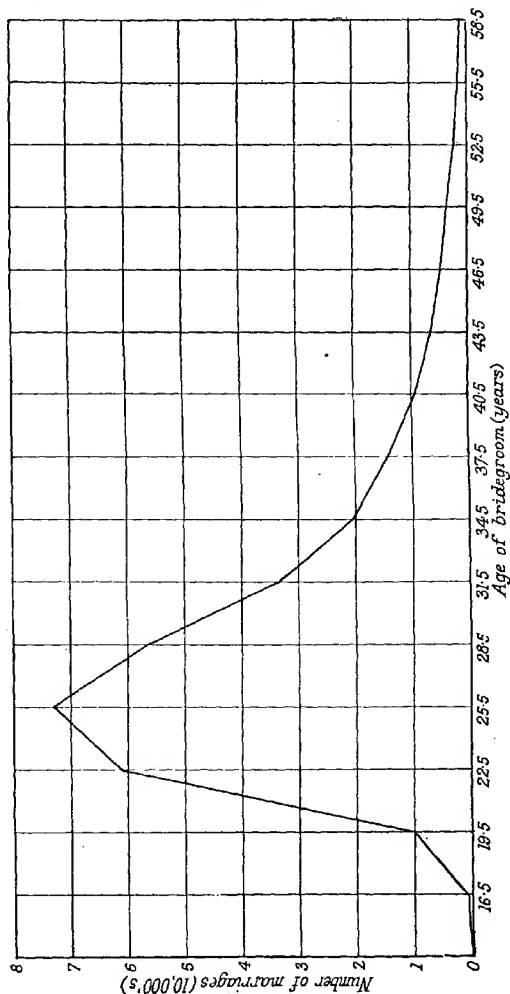
FIG. 6.8.—Frequency-distribution of Australian Marriages, classified according to the Bridegroom's Age.   (Table 6.8.)

we may explain at this stage that the skewness is said to be positive if the longer tail of the curve lies to the right, or negative if it lies to the

7

left ; *e.g.* the curve of fig. 6.8 has positive skewness, whilst those of figs. 6.9 and 6.10 have negative skewness.

TABLE 6.9.—*Showing the Frequency-distribution of Fecundity, i.e. the Ratio of the Number of Yearling Foals Produced to the Number of Coverings, for Brood-mares (Race-horses) Covered Eight Times at Least.* (Pearson, Lee and Moore, *Phil. Trans.*, A, vol. 192, 1899, p. 303.) (See fig. 6.9.)

Fecundity.	Number of Mares with Fecundity between the Given Limits.	Fecundity.	Number of Mares with Fecundity between the Given Limits.
1/30– 3/30	2	17/30–19/30	315
3/30– 5/30	7·5	19/30–21/30	337
5/30– 7/30	11·5	21/30–23/30	293·5
7/30– 9/30	21·5	23/30–25/30	204
9/30–11/30	55	25/30–27/30	127
11/30–13/30	104·5	27/30–29/30	49
13/30–15/30	182	29/30–1	19
15/30–17/30	271·5		
		Total	2000·0
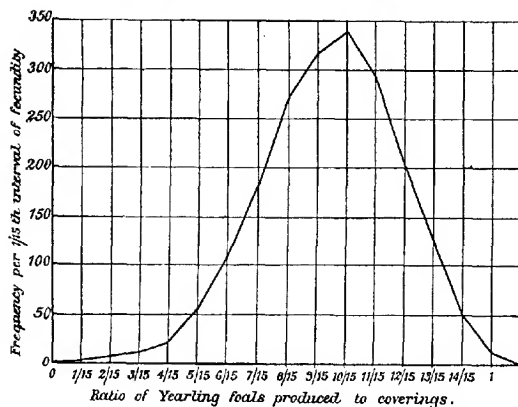
FIG. 6.9.—Frequency-distribution of Fecundity for Brood-mares. (Table 6.9.)

The Extremely Asymmetrical, or J-shaped, Distribution.

6.21. In this type the class-frequencies run up to a maximum at one end of the range, as in fig. 6.12.

This may be regarded as a limiting form of the previous distribution, and, in fact, the two cannot always be distinguished by elementary methods if the original data are not available. If, for instance, the frequencies of Table 6.11 had been given by five-year intervals only, they would have run 322, 213, 70, 27, etc., thus suggesting that the maximum number of deaths

TABLE 6.10.—*Showing Barometric Heights at Greenwich on Alternate Days from 1848–1926.* (Data from S. J. Pretorius, "Skew Bivariate Frequency Surfaces," *Biometrika*, vol. 22, 1930–31, p. 154.)   (See fig. 6.10.)

Barometric Height (Central Value in Inches).	Number of Days.	Barometric Height (Central Value in Inches).	Number of Days.
28·35	1	29·65	3176
28·45	4	29·75	3700
28·55	12	29·85	3921
28·65	43	29·95	3749
28·75	60	30·05	2951
28·85	81	30·15	1951
28·95	189	30·25	1148
29·05	282	30·35	563
29·15	542	30·45	258
29·25	813	30·55	73
29·35	1233	30·65	13
29·45	1752	30·75	7
29·55	2333		
		Total	28,855
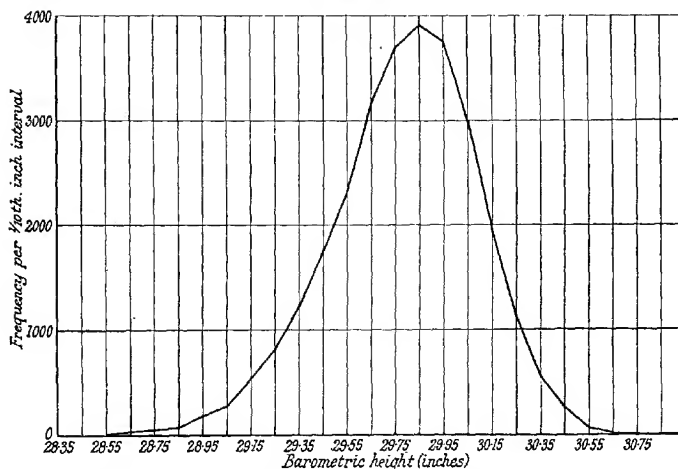
FIG. 6.10.—Barometric Height at Greenwich on Alternate Days from 1848–1926.   (Table 6.10.)

occurred at the beginning of life, *i.e.* that the distribution was J-shaped. It is only the analysis of deaths in the earlier years by one-year intervals which shows that the frequencies reach a maximum in the third year and that therefore the distribution is of the moderately asymmetrical type.

TABLE 6.11.—*Showing the Number of Deaths from Scarlet Fever at Different Ages in England and Wales in 1933.* (Data from Registrar-General's Statistical Review of England and Wales for 1933, Tables, Part I, Medical, supplemented by information supplied by him in correspondence.)   (See fig. 6.11.)

Age in Years.	Number of Deaths.	Number per Year.
0–	16	16
1–	69	69
2–	89	89
3–	74	74
4–	74	74
5–	213	42·6
10–	70	14·0
15–	27	5·4
20–	26	5·2
25–	17	3·4
30–	12	2·4
35–	11	2·2
40–	10	2·0
45–	6	1·2
50–	7	1·4
55–	5	1·0
60–	—	—
65–	1	0·2
70–	1	0·2
75–	1	0·2
80–	—	—
Total	729	—

In practical cases no hard-and-fast rule can be drawn between the moderately and extremely asymmetrical types, any more than between the asymmetrical and the symmetrical types.

6.22. In economic statistics this form of distribution is particularly characteristic of the distribution of wealth in the population at large, as illustrated by income tax and house valuation returns, and the curve to which it gives rise has been called the " Pareto line," after Vilfredo Pareto, who directed the attention of economists to it (*vide* ref. (99)).   The student should draw the histogram of the data of Table 6.5 in illustration of this point.

Such distributions may, of course, be a very extreme case of the last type.   It is difficult to say.   But if the maximum is not absolutely at the lower end of the range, it is very close thereto.

Official returns do not usually give the necessary analysis of the frequencies at the lower end of the range to enable the exact position of the maximum to be determined ; and for this reason the data on which Table 6.12 is founded, though of course very unreliable, are of some interest.   It will be seen from the table and fig. 6.13 that with the given classification the distribution appears clearly assignable to the present type, the number of estates between zero and £100 in annual value being more than six times as great as the number between £100 and £200 in annual value, and the frequency continuously falling as the value increases.   A close analysis of the first class suggests, however, that the greatest frequency does not occur

actually at zero, but that there is a true maximum frequency for estates of about £1 15/- in annual value. The distribution might therefore be more correctly assigned to the second type, but the position of the greatest frequency indicates a degree of skewness which is high even compared with the skewness of fig. 6.11.

The type is not very frequent in other classes of material, but instances occur here and there. Distributions of deaths of centenarians afford an
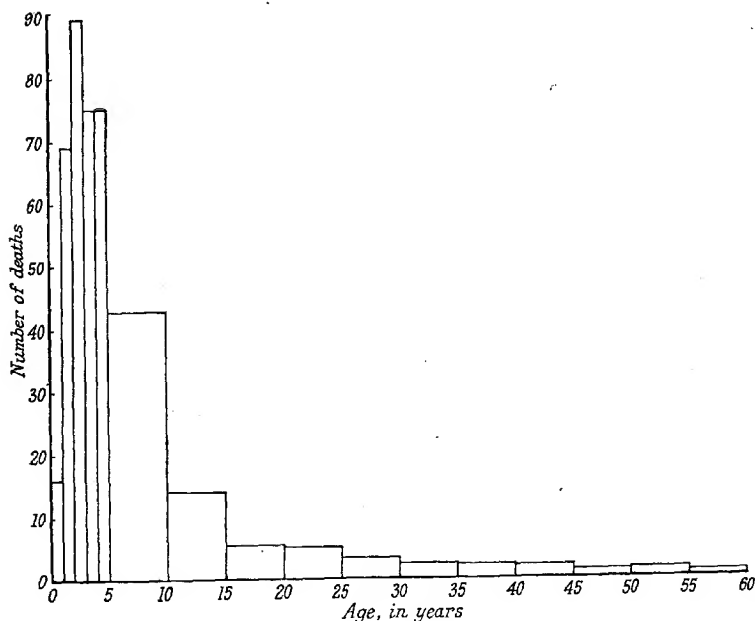
FIG. 6.11.—Histogram of Number of Deaths from Scarlet Fever for Various Ages. (Table 6.11.)

example, and so, curiously enough, do deaths of infants unless the class-interval is exceedingly fine—a matter of hours. It has also been shown that the distribution may be obtained by compiling the frequencies of the numbers of genera with 1, 2, 3, . . . species in any biological group. Table 6.13 shows such a distribution for the Chrysomelid beetles.

The U-shaped Distribution.

6.23. This type exhibits a maximum frequency at the ends of the range and a minimum towards the centre, as in fig. 6.14.

This is a rare but interesting form of distribution, as it stands in somewhat marked contrast to the preceding forms. Table 6.14 and fig. 6.15

illustrate an example based on a considerable number of observations, viz. the distribution of degrees of cloudiness, or estimated percentage of the sky covered by cloud, at Greenwich in July.

For the purposes of the illustration we regard cloudiness as a variate varying from complete overcastness to clear sky, the range being divided into eleven *equal* parts.

It will be seen that a sky completely or almost completely overcast at
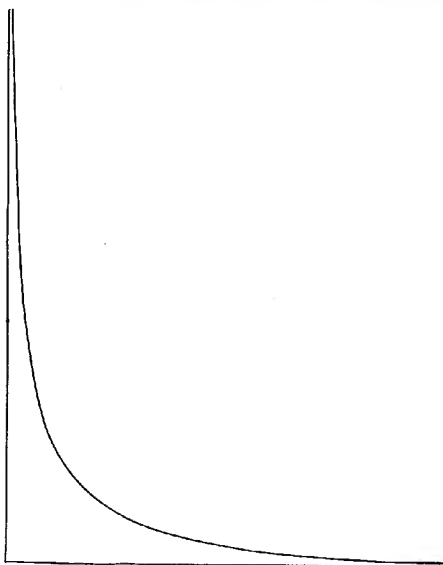
FIG. 6.12.—An ideal Distribution of the Extremely Asymmetrical Form.

the time of observation is the most common, a practically clear sky comes next, and the intermediates are more rare.

The remarks we made about the extreme end of the J-shaped distribution also apply to the U-shaped distribution. In particular cases it may be that the grouping is too coarse to reveal the true character of the frequency at the maxima, and if the data were more complete we might discover that the two arms of the U in fact were bent over.

Truncated Forms.

6.24. The four types we have been considering sometimes occur in an incomplete form. Certain limitations on the range of the variate may result in a kind of truncation at one end or the other. Consider, for example, Table 6.15, p. 107. In obtaining these figures, twelve dice were thrown and the occurrence of a 6 was called a success. At one throw there could thus be any number of successes from 0 to 12. The dice were thrown 4096 times.
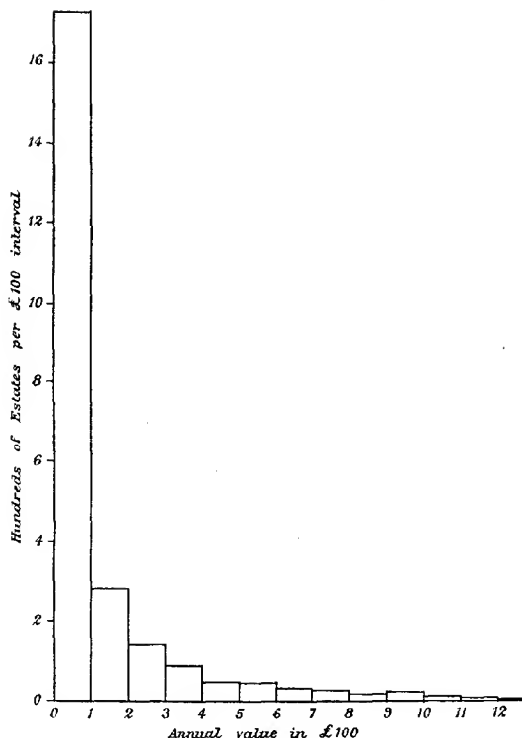
FIG. 6.13.—Frequency-distribution of the Annual Values of certain Estates
in England in 1715; 2476 Estates. (Table 6.12.)

Fig. 6.16 gives the frequency-polygon for this distribution. We can picture it as a slightly skew distribution which has been cut off on the left owing to the inadmissibility of negative values of the variate. Discontinuous variates not infrequently give rise to this effect of truncation.

Complex Distributions.

6.25. Table 6.16 gives the number of male deaths within certain age-limits for England and Wales in the years 1930–32.

The histogram for these data is given in fig. 6.17. It will be seen that the distribution has three maxima, one for each of the 0–5, the 20–25 and the 70–75 age-groups.

Without looking too closely into this mortality curve we can see that the high frequency at the beginning is undoubtedly due to the heavy infantile death-rate. We can, if we choose, regard the distribution as
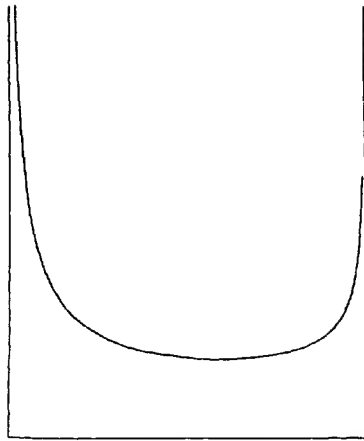
Fig. 6.14.—An Ideal Distribution of the **U**-shaped Form.
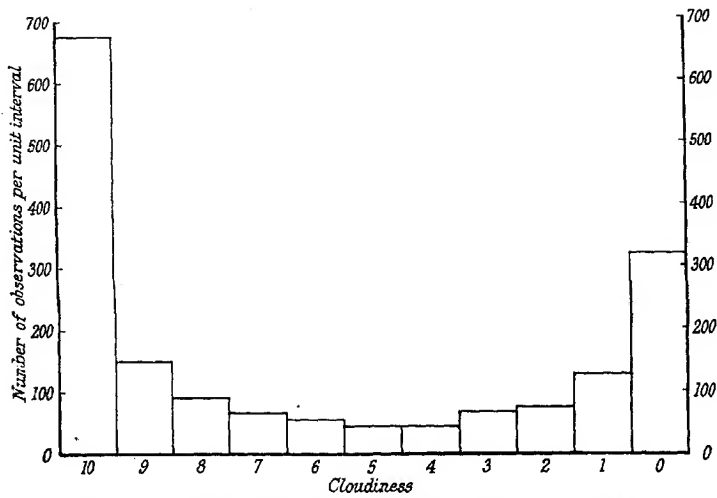
Fig. 6.15.—Cloudiness at Greenwich in July; 1715 Observations.   (Table 6.14.)

TABLE 6.12.—*Showing the Numbers and Annual Values of the Estates of those who had taken part in the Jacobite Rising of 1715.* (Compiled from Cosin's "*Names of the Roman Catholics, Nonjurors, and others who Refused to take the Oaths to his late Majesty King George, etc.*"; London, 1745. Figures of very doubtful absolute value. See a note in Southey's "*Commonplace Book,*" vol. 1, p. 573, quoted from the Memoirs of T. Hollis.) (See fig. 6.13.)

Annual Value in £100.	Number of Estates.	Annual Value in £100.	Number of Estates.
0– 1	1726·5	17–18	1
1– 2	280	—	—
2– 3	140·5	20–21	4
3– 4	87	21–22	1
4– 5	46·5	22–23	1
5– 6	42·5	23–24	1
6– 7	29·5	—	—
7– 8	25·5	27–28	2
8– 9	18·5	—	—
9–10	21	31–32	1
10–11	11·5	—	—
11–12	9·5	39–40	1
12–13	4	—	—
13–14	3·5	45–46	1
14–15	8	—	—
15–16	3	48–49	1
16–17	5		
		Total	2476

made up by the superposition of three others : a J-shaped distribution for the lower years, a small one-humped distribution with its maximum about the period 20–25 years, and a skew distribution for the higher ages. This is an example of the fact we have already mentioned, that a complex distribution can sometimes be analysed into simpler types. In this particular case the analysis is likely to be of real service in actuarial work and in investigations into the causes of death.

6.26. Finally, we give an example of a pseudo-frequency-distribution of a type occasionally resorted to when the data can be classified according to a characteristic which, though not strictly speaking measurable, can nevertheless be graduated in an ordered sequence. Such a case arises fairly often in psychological work.

A list of 100 words was read out to each of 11 subjects. Subsequently, at 15-minute intervals, four fresh lists were read out which contained 25 of the words in the original and 25 new words, the four taken together accounting for the whole of the original 100. The subject had to say whether these individual words were in the original list or not, and to state whether he was certain, fairly sure, doubtful but inclined one way or the other, or merely doubtful. The various phases of belief were then allotted numbers, and ran from −3 (certainty that a word was not in the original) through 0 (doubt, without inclination one way or the other) to +3 (certainty that a word was in the original). The tabulation on p. 108 sets out the results for words in the original list (data reproduced by permission from the records of the Department of Psychology, University of St Andrews).

TABLE 6.13.—Chrysomelidæ (*beetles*).   *Numbers of Genera with 1, 2, 3, . . . Species.*
(Compiled by Dr J. C. Willis, F.R.S.; cited from G. U. Yule, "A Mathematical
Theory of Evolution based on the Conclusions of Dr J. C. Willis," *Phil. Trans.*,
B, vol. 213, 1924, p. 85.)

Species.	Genera.	Species.	Genera.	Species.	Genera.
1	215	32	1	74	1
2	90	33	1	76	1
3	38	34	1	77	1
4	35	35	1	79	1
5	21	36	3	83	1
6	16	37	1	84	3
7	15	38	1	87	2
8	14	39	2	89	1
9	5	40	2	92	2
10	15	41	1	93	1
11	8	43	4	110	1
12	9	44	1	114	1
13	5	45	1	115	1
14	6	46	1	128	1
15	8	49	2	132	1
16	6	50	4	133	1
17	6	52	1	146	1
18	3	53	1	163	1
19	4	56	1	196	1
20	3	58	1	217	1
21	4	59	1	227	1
22	4	62	1	264	1
23	5	63	3	327	1
24	4	65	1	399	1
25	2	66	1	417	1
26	3	67	1	681	1
27	1	69	1		
28	3	71	1		
29	3	72	1	Total	627
30	3	73	1		

TABLE 6.14.—*Showing the Frequencies of Estimated Intensities of Cloudiness at Greenwich
during the Years 1890–1904 (excluding 1901) for the Month of July.*   (Data from
Gertrude E. Pearse, *Biometrika*, vol. 20A, 1928, p. 336.)   (See fig. 6.15.)

Degrees of Cloudiness.	Frequency.	Degrees of Cloudiness.	Frequency.
10	676	4	45
9	148	3	68
8	90	2	74
7	65	1	129
6	55	0	320
5	45		
		Total	1715

TABLE 6.15.—*Twelve Dice thrown 4096 Times, a Throw of 6 Points reckoned as a Success* (Weldon's data; cited by F. Y. Edgeworth, *Encyclopedia Britannica*, 11th ed., vol. 22, p. 39.) (See fig. 6.16.)

Number of Successes .	0	1	2	3	4	5	6	7 and over	Total.
Number of Throws .	447	1145	1181	796	380	115	24	8	4096
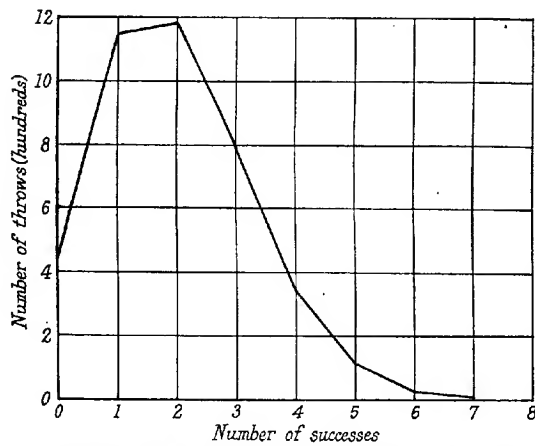
FIG. 6.16.—Frequency Polygon of Successes with Dice Throwing. (Table 6.15.)

TABLE 6.16.—*Showing the Number of Male Deaths in England and Wales for 1930–32, classified by Ages at Death.* (Data from Registrar-General's Statistical Review of England and Wales, 1933, Text.) (See fig. 6.17.)

Age at Death (years).	Number of Deaths.	Age at Death (years).	Number of Deaths.
0– 5	97,290	55– 60	56,639
5–10	11,532	60– 65	68,103
10–15	7,305	65– 70	80,690
15–20	13,062	70– 75	84,041
20–25	16,741	75– 80	72,180
25–30	16,126	80– 85	45,094
30–35	15,673	85– 90	19,913
35–40	18,345	90– 95	5,145
40–45	23,778	95–100	767
45–50	33,158	100 and over	48
50–55	43,812		
		Total	729,442

Words in the original list were classified as :

In			Possibly either In or Out.	Out.		
Certain.	Fairly Sure.	Doubtful.		Doubtful.	Fairly Sure.	Certain.
+3	+2	+1	0	−1	−2	−3
540	117	63	39	63	87	191

These results are very curious, and are borne out by other data of a similar kind. In particular we see that there were more cases of certainty about something which was not true than of doubt without inclination.
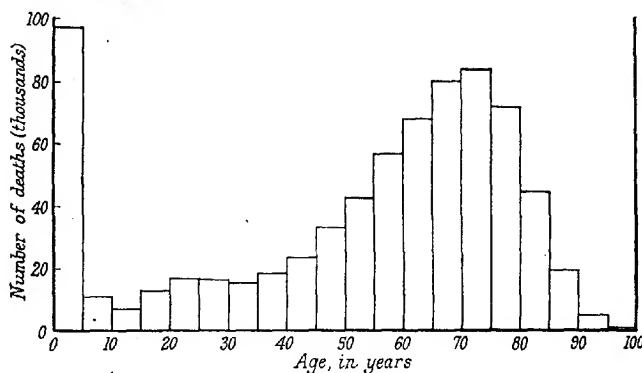
FIG. 6.17.—Histogram of Number of Deaths at Various Ages.   (Table 6.16.)

In this example we are clearly making some assumption in allotting numbers to various degrees of belief; but it would be impossible to measure belief on a scale, and we have to do the best we can.   The numbers attached to the variate in such cases are not measures, but convenient ordinals, like the numbers attached to kings of the same name.   For this reason a frequency diagram of such data can only give a very general idea of their true nature.

SUMMARY.

1. Data in which the individuals are specified by the numerical values of a variable, or variate, may with convenience be arranged in a table which gives the frequency lying within successive, preferably equal, ranges of the variable.   Such an arrangement is called a frequency-distribution.

2. The frequency-distribution can be represented diagrammatically by means of a frequency-polygon or a histogram.

3. The histogram is particularly appropriate to cases in which the frequency changes rapidly or the class-intervals are not all of the same width.

4. As the width of the class-intervals becomes smaller, the frequency-polygon or the histogram may be imagined to approach a smooth curve, which is called the frequency-curve.

5. A large number of frequency distributions occurring in practice fall into four types : the symmetrical, the moderately asymmetrical or skew, the extremely asymmetrical or J-shaped and the U-shaped types. Certain other distributions can be analysed into constituents each of which belongs to one of these types.

EXERCISES.

6.1. If the diagram fig. 6.6 is redrawn to scales of 300 observations per interval to the inch and 4 inches of stature to the inch, what is the scale of observations to the square inch?

If the scales are 100 observations per interval to the centimetre and 2 inches of stature to the centimetre, what is the scale of observations to the square centimetre?

6.2. If fig. 6.10 is redrawn to scales of 900 days to the inch and 0·3 inch of barometric height to the inch, what is the scale of observations to the square inch?

If the scales are 400 days to the centimetre and 0·1 inch of barometric height to the centimetre, what is the scale of observations to the square centimetre?

6.3. If a frequency-polygon be drawn to represent the data of Table 6.1, what number of observations will the polygon show between birth-rates of 16·5 and 17·5 per thousand, instead of the true number 89?

6.4. If a frequency-polygon be drawn to represent the data of Table 6.6, what number of observations will the polygon show between head-breadths 5·95 and 6·05, instead of the true number 236?

6.5. Draw frequency-polygons or histograms, as the case seems to require, for the following distributions, and assign them to the four types we have enumerated in 6.18 :—

(a) *Size of Firms in the Food, Drink and Tobacco Trades of Great Britain.* (Final Report of the Fourth Census of Production, 1930, Part III.) The following table shows the number of firms employing on an average certain numbers of persons :—

Size of Firm (Average Numbers Employed).	11–24	25–49	50–99	100–199	200–299	300–399	400–499	500–749	750–999	1000–1499	1500 and over	Total
Number of Firms .	2245	1449	771	439	164	75	36	54	31	23	29	5316

(b) *The Percentages of Deaf-mutes among Children of Parents One of whom at least was a Deaf-mute, for Marriages producing Five Children or More.* (Compiled from material in "*Marriages of the Deaf in America*," ed. E. A. Fay, Volta Bureau, Washington, 1898.)

Percentage of Deaf-mutes.	Number of Families.	Percentage of Deaf-mutes.	Number of Families.
0–20	220	60– 80	5·5
20–40	20·5	80–100	15
40–60	12	Total	273

(c) *Yield of Grain in pounds from Plots of $\frac{1}{500}$th Acre in a Wheat Field.* (Mercer and Hall, "The Experimental Error of Field Trials," *Journ. Agr. Science*, vol. 4, 1911, p. 107.)

Yield of Grain in pounds per $\frac{1}{500}$th Acre. (Central value of range.)	2·8	3·0	3·2	3·4	3·6	3·8	4·0	4·2	4·4	4·6	4·8	5·0	5·2	Total
Number of Plots. .	4	15	20	47	63	78	88	69	59	35	10	8	4	500

(d) *The Frequencies of Different Numbers of Petals for Three Series of Ranunculus bulbosus.* (H. de Vries, *Ber. deutsch. bot. Ges.*, Bd. 12, 1894, *q.v.* for details.)

Number of Petals.	Frequency.		
	Series A.	Series B.	Series C.
5	312	345	133
6	17	24	55
7	4	7	23
8	2	—	7
9	2	2	2
10	—	—	2
11	—	2	—
Total	337	380	222

6.6. A number of perfectly spherical balls, all of the same material, give a symmetrical distribution when classified according to their diameters. Show that, if they are classified according to their weights, their frequency-distribution will be positively skew towards the higher weights.

In the light of this result compare the distributions of Table 6.7 with the distributions of the table on p. 111.

6.7. Toss a coin six times and note the number of heads. Repeat the experiment 100 times or more, and draw a frequency-polygon of your results classified according to the number of heads at each throw.

6.8. Find the frequency-distribution of 200 bars of a waltz by Strauss classified according to the number of notes in the treble clef of each bar, and compare it with a similar distribution from modern waltzes.

6.9. Examine qualitatively the effect on the distribution of Table 6.8 of an allowance for the fact that minors tend to overstate their age when marrying.

6.10. The distribution of a herd of cows classified according to the quantity of milk produced by each cow per week is symmetrical. The distribution of the same herd classified according to the amount of butter-fat produced by each cow per week is negatively skew towards the lower quantities. Suggest a possible explanation for this fact.

The Frequency-distribution of Weights for Adult Males born in England, Scotland, Wales and Ireland. (Loc. cit., Table 6.7.) Weights were taken to the nearest pound, consequently the true Class-intervals are 89·5-99·5, 99·5-109·5, etc.

Weight in lbs.	Number of Men within given Limits of Weight. Place of Birth—				Total.
	England.	Scotland.	Wales.	Ireland.	
90–	2	—	—	—	2
100–	26	1	2	5	34
110–	133	8	10	1	152
120–	338	22	23	7	390
130–	694	63	68	42	867
140 -	1240	173	153	57	1623
150–	1075	255	178	51	1559
160–	881	275	134	36	1326
170–	492	168	102	25	787
180–	304	125	34	13	476
190–	174	67	14	8	263
200–	75	24	7	1	107
210–	62	14	8	1	85
220 -	33	7	1	—	41
230–	10	4	2	—	16
240–	9	2	—	—	11
250–	3	4	1	—	8
260–	1	—	—	—	1
270–	—	—	—	—	—
280–	—	—	1	—	1
Total	5552	1212	738	247	7749

CHAPTER 7.

AVERAGES AND OTHER MEASURES OF LOCATION.

The Principal Characteristics of Frequency-distributions.

7.1. The condensation of data into a frequency-distribution is a first and necessary step in rendering a long series of observations comprehensible. But for practical purposes it is not enough, particularly when we want to compare two or more different series. As a next step we wish to be able to define quantitatively the characteristics of a frequency-distribution in as few numbers as possible.

7.2. It might seem at first sight that very difficult cases of comparison of two distributions could arise in which, for example, we had to contrast a symmetrical distribution with a J-shaped distribution. In practice, however, we rarely have to deal with such a case. Distributions drawn from similar material are usually of similar form—as, for instance, when we wish to compare the distributions of stature in two races of man, or the birth-rates in English registration districts in two successive decades, or the numbers of wealthy people in two different countries. The practical use of the various statistical quantities which we shall discuss in this and the next two chapters is based on this fact.

7.3. There are two fundamental characteristics in which similar frequency-distributions may differ :

(1) They may differ markedly in position, *i.e.* in the value of the variate round which they centre, as in fig. 7.1, A.

(2) They may differ in the extent to which the observations are dispersed about the central value. Figs. 7.1, B and C, show cases in which distributions differ in dispersion only, and in both dispersion and position, respectively.

To these two characteristics we may add a third group of less importance, comprising differences in skewness, peakedness, and so on.

Measures of the first character, *i.e.* position or location, are generally known as **averages**. Measures of the second are termed **measures of dispersion**. Measures of the properties in the third group have each their appropriate name, which we shall give when we come to consider them in detail.

The present chapter deals only with averages. Chapter 8 deals with measures of dispersion, whilst Chapter 9 deals with the remaining quantities.

Dimensions of an Average.

7.4. In whatever way an average is defined, it may be as well to note it is merely a certain value of the variable, and is therefore necessarily of the same *dimensions* as the variable : *i.e.* if the variable be a

112

length, its average is a length ; if the variable be a percentage, its average is a percentage ; and so on.   But there are several different ways of approximately defining the position of a frequency-distribution—that is,
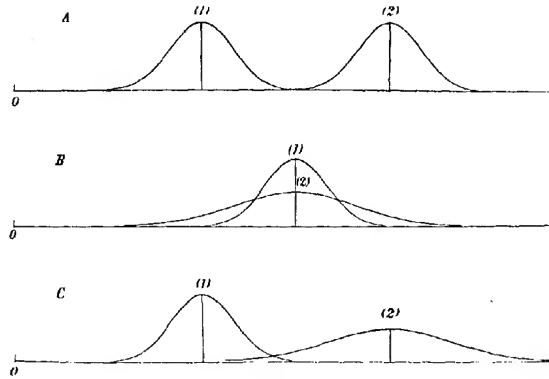
Fig. 7.1.

there are several different forms of average, and the question therefore arises, By what criteria are we to judge the relative merits of different forms ?   What are, in fact, the desirable properties for an average to possess ?

Desiderata for a Satisfactory Average.

7.5.   (*a*) In the first place, it almost goes without saying that an average should be rigidly defined, and not left to the mere estimation of the observer.   An average that was merely estimated would depend too largely on the observer as well as the data.

(*b*) An average should be based on all the observations made.   If not, it is not really a characteristic of the whole distribution.

(*c*) It is desirable that the average should possess some simple and obvious properties to render its general nature readily comprehensible : an average should not be of too abstract a mathematical character.

(*d*) It is, of course, desirable that an average should be calculated with reasonable ease and rapidity.   Other things being equal, the easier calculated is the better of two forms of average.   At the same time great weight must not be attached to mere ease of calculation, to the neglect of other factors.

(*e*) It is desirable that the average should be as little affected as may be possible by what we have termed *fluctuations of sampling*.   If different samples be drawn from the same material, however carefully they may be taken, the averages of the different samples will rarely be quite the same, but one form of average may show much greater differences than another.   Of the two forms, the more stable is the better.   The full discussion of this condition must, however, be postponed to a later section of this work (Chap. 20).

8

(*f*) Finally, by far the most important desideratum is this, that the measure chosen shall lend itself readily to algebraical treatment. If, *e.g.*, two or more series of observations on similar material are given, the average of the combined series should be readily expressed in terms of the averages of the component series; if a variable may be expressed as the sum of two or more others, the average of the whole should be readily expressed in terms of the averages of its parts. A measure for which simple relations of this kind cannot be readily determined is likely to prove of somewhat limited application.

7.6. There are three forms of average in common use, the **arithmetic mean**, the **median** and the **mode**, the first named being by far the most widely used in general statistical work. To these may be added the **geometric mean** and the **harmonic mean**, more rarely used, but of service in special cases. We will consider these in the order named.

The Arithmetic Mean.

7.7. The arithmetic mean of a series of values of a variable X_1, X_2, X_3, . . . X_N, N in number, is the quotient of the sum of the values by their number. That is to say, if M be the arithmetic mean,

$$M = \frac{1}{N}(X_1 + X_2 + X_3 + \ . \ . \ . \ + X_N)$$

The arithmetic mean is also denoted by placing a bar over the variate symbol, so that we may also write :

$$\bar{X} = \frac{1}{N}(X_1 + X_2 + \ . \ . \ . \ + X_N)$$

To express these formulæ more briefly by the use of the summation symbol S,

$$\bar{X} = M = \frac{1}{N}S(X) \qquad . \qquad . \qquad . \qquad . \quad (7.1)$$

The word *mean* or *average* alone, without qualification, is very generally used to denote this particular form of average ; that is to say, when anyone speaks of " the mean " or " the average " of a series of observations, it may, as a rule, be assumed that the arithmetic mean is meant.

7.8. It is evident that the arithmetic mean fulfils the conditions laid down in (*a*) and (*b*) of **7.5**, for it is rigidly defined and based on all the observations made. Further, it fulfils condition (*c*), for its general nature is readily comprehensible. If the wages-bill for N workmen is £P, the arithmetic mean wage, P/N pounds, is the amount that each would receive if the whole sum available were divided equally between them : conversely, if we are told that the mean wage is £M, we know this means that the wages-bill is NM pounds. Similarly, if N families possess a total of C children, the mean number of children per family is C/N—the number that each family would possess if the children were shared uniformly. Conversely, if the mean number of children per family is M, the total number of children in N families is NM. The arithmetic mean expresses, in fact, a simple relation between the whole and its parts.

The mean is also satisfactory as regards conditions (*e*) and (*f*), but we shall have to defer proof of this statement for the present.

Calculation of the Arithmetic Mean.

7.9. As regards condition (d), simplicity of calculation, the mean takes a high place. In the cases just cited, it will be noted that the mean is actually determined without even the necessity of determining or noting all the individual values of the variable : to get the mean wage we need not know the wages of every hand, but only the wages-bill; to get the mean number of children per family we need not know the number in each family, but only the total. If this total is not given, but we have to deal with a moderate number of observations —so few (say 30 or 40) that it is hardly worth while compiling the frequency-distribution—the arithmetic mean is calculated directly as suggested by the definition, *i.e.* all the values observed are added together and the total divided by the number of observations.

7.10. But if the number of observations be large, the process of adding together all the values of the variate may be prohibitively lengthy. It may be shortened considerably by forming the frequency-table and treating all the values in each class as if they were identical with the mid-value of the class-interval, a process which in general gives an approximation that is quite sufficiently exact for practical purposes if the class-interval has been taken moderately small. In this process each class-frequency is multiplied by the mid-value of the interval, the products added together, and the total divided by the number of observations. If f denote the frequency of any class, X the mid-value of the corresponding class-interval, the value of the mean so obtained may be written :

$$M = \frac{1}{N}S(fX) \qquad . \qquad . \qquad . \qquad . \qquad (7.2)$$

7.11. But this procedure is still further abbreviated in practice by the following artifices : (1) The class-interval is treated as the unit of measurement throughout the arithmetic ; (2) the difference between the mean and the mid-value of some arbitrarily chosen class-interval is computed instead of the absolute value of the mean.

If A be the arbitrarily chosen value and

$$X = A + \xi \quad . \qquad . \qquad . \qquad . \qquad (7.3)$$

then

$$S(fX) = S(fA) + S(f\xi)$$

or, since A is a constant,

$$M = A + \frac{1}{N}S(f\xi) \qquad . \qquad . \qquad . \qquad (7.4)$$

The calculation of $S(fX)$ is therefore replaced by the calculation of $S(f\xi)$. The advantage of this is that the class-frequencies need only be multiplied by small integral numbers ; for A being the mid-value of a class-interval, and X the mid-value of another, and the class-interval being treated as a unit, the ξ's must be a series of integers proceeding from zero at the arbitrary origin A. To keep the values of ξ as small as possible, A should be chosen near the middle of the range.

It may be mentioned here that $\frac{1}{N}S(\xi)$, or $\frac{1}{N}S(f\xi)$ for the grouped

distribution, is sometimes termed the **first moment** of the distribution about the arbitrary origin A.

Example 7.1.—As an example, let us find the arithmetic mean of the heights in the distribution of Table 6.7. In this case the class-interval is a unit (1 inch), so the value of $M - A$ is given directly by dividing $S(f\xi)$ by N. The student must notice that, measures having been made to the nearest eighth of an inch, the mid-values of the intervals are $57\frac{7}{16}$, $58\frac{7}{16}$, etc., and not 57·5, 58·5, etc.

CALCULATION OF THE MEAN: *Calculation of the Arithmetic Mean Stature of Male Adults in the British Isles from the Figures of Table 6.7, p. 94.*

(1)	(2)	(3)	(4)
Height, Inches.	Frequency f.	Deviation from Arbitrary Value A ξ.	Product $f\xi$.
57-	2	− 10	20
58-	4	− 9	36
59-	14	− 8	112
60-	41	− 7	287
61-	83	− 6	498
62-	169	− 5	845
63-	394	− 4	1576
64-	669	− 3	2007
65-	990	− 2	1980
66-	1223	− 1	1223
67-	1329	0	− 8584
68-	1230	+ 1	1230
69-	1063	+ 2	2126
70-	646	+ 3	1938
71-	392	+ 4	1568
72-	202	+ 5	1010
73-	79	+ 6	474
74-	32	+ 7	224
75-	16	+ 8	128
76-	5	+ 9	45
77-	2	+10	20
Total	8585	−	+8763

$$S(f\xi) = +8763 - 8584 = +179$$

$$M - A = +\frac{179}{8585} = +0.02 \text{ class-intervals or inches.}$$

$$\therefore \quad M = 67\frac{7}{16} + 0.02 = 67.46 \text{ inches.}$$

7.12. As calculations of the mean constantly have to be made, the student should familiarise himself with the process we have just illustrated, and note that a check can always be effected on the arithmetic in the following way :—

Since
$$f(\xi + 1) = f\xi + f$$
$$S\{f(\xi + 1)\} = S(f\xi) + S(f)$$
$$S\{f(\xi + 1)\} - S(f\xi) = S(f)$$
$$= \text{Total frequency}$$

Hence, if we tabulate the values of $f(\xi + 1)$ as well as those of $f\xi$ and find their totals, the difference must, if the arithmetic is correct, be equal to the total frequency.

7.13.   It will be evident that a classification by unequal intervals is, at best, a hindrance in the calculation of the mean, and the use of an indefinite interval at the end of the distribution renders exact calculation impossible.   The following example illustrates the calculation for unequal class-intervals and the arithmetical check to which we have just referred.

Example 7.2.—Data from Table 6.11, page 100.   What is the average age at death from scarlet fever ?

Here there is a change of the class-interval at the five-year point.   We take a year to be the unit, and the centre of the interval 5–10 years as an arbitrary origin, which means that $A = 7\cdot5$ years.

CALCULATION OF THE MEAN: *Calculation of the Arithmetic Mean Age of Persons Dying from Scarlet Fever in the United Kingdom in 1933 (Table 6.11, p. 100)*

Age, Years.	Frequency, f.	Deviation from A, ξ.	$f\xi$.	$f(\xi + 1)$.
0–	16	−7	− 112	− 96
1–	69	−6	− 414	− 345
2–	89	−5	445	356
3–	74	−4	− 296	− 222
4–	74	−3	− 222	− 148
5–	213	0	− 1489	− 1167
				213
10–	70	5	350	420
15–	27	10	270	297
20–	26	15	390	416
25–	17	20	340	357
30–	12	25	300	312
35–	11	30	330	341
40–	10	35	350	360
45–	6	40	240	246
50–	7	45	315	322
55–	5	50	250	255
60–	—	55	—	—
65–	1	60	60	61
70–	1	65	65	66
75–	1	70	70	71
Total	729	—	+ 3330	+ 3737

Hence,
$$S(f\xi) = 3330 - 1489 = 1841$$
and
$$S\{f(\xi + 1)\} = 3737 - 1167 = 2570$$
and the difference $2570 - 1841 = 729$, as it should.
Hence,
$$M - A = \frac{1841}{729} = 2\cdot525 \text{ years}$$
and
$$M = 7\cdot5 + 2\cdot525 = 10\cdot025 \text{ years}$$

7.14. We return again below, in **7.16** (*c*), to the question of the errors caused by the assumption that all values within the same interval may be treated as approximately the mid-value of the interval. It is sufficient to say here that the error is in general very small and of uncertain sign for a distribution of the symmetrical or only moderately asymmetrical type, provided, of course, the class-interval is not large. In the case of the " J-shaped " or extremely asymmetrical distribution, however, the error is evidently of definite sign, for in all the intervals the frequency is piled up at the limit lying towards the greatest frequency, *i.e.* the lower end of the range in the case of the illustrations given in Chapter 6, and is not evenly distributed over the interval. In distributions of such a type the intervals must be made very small indeed to secure an approximately accurate value for the mean. The student should test for himself the effect of different groupings in two or three different cases, so as to get some idea of the degree of inaccuracy to be expected.

7.15. If a diagram has been drawn representing the frequency-distribution, the position of the mean may conveniently be indicated by a
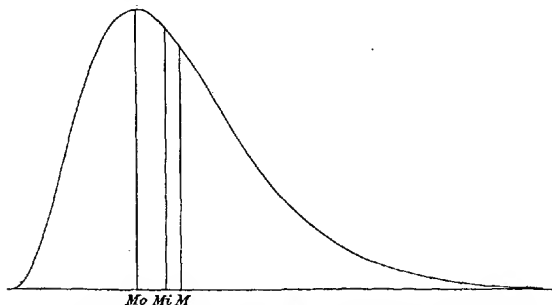
FIG. 7.2.—Mean *M*, Median *Mi* and Mode *Mo* of the Ideal Moderately
Asymmetrical Distribution.

vertical through the corresponding point on the base. In a moderately asymmetrical distribution the mean lies on the side of the greatest frequency towards the longer " tail " of the distribution : *M* in fig. 7.2 shows the position of the mean in an ideal distribution. In a symmetrical distribution the mean coincides with the centre of symmetry. The student should mark the position of the mean in the diagram of every frequency-distribution that he draws, and so accustom himself to thinking of the mean not as an abstraction, but always in relation to the frequency-distribution of the variable concerned.

Properties of the Arithmetic Mean.

7.16. The following are important properties of the arithmetic mean, and the examples illustrate the facility of its algebraic treatment :—

(*a*) The sum of the deviations from the mean, taken with their proper signs, is zero.

This follows at once from equation (7.4) : for if *M* and *A* are identical, evidently $S(f\xi)$ must be zero.

(b) If a series of N observations of a variable X consist of, say, two component series, the mean of the whole series can be readily expressed in terms of the means of the two components. For if we denote the values in the first series by X_1 and in the second series by X_2,

$$S(X) = S(X_1) + S(X_2)$$

that is, if there be N_1 observations in the first series and N_2 in the second, and the means of the two series be M_1, M_2, respectively,

$$NM = N_1 M_1 + N_2 M_2 \quad . \qquad . \qquad . \qquad . \quad (7.5)$$

For example, we find from the data of Table 6.7,

Mean stature of the 346 men born in Ireland = 67·78 inches
,,   ,,   ,,   741   ,,   ,,   Wales = 66·62   ,,

Hence the mean stature of the 1087 men born in the two countries is given by the equation

$$1087M = (346 \times 67·78) + (741 \times 66·62)$$

that is, $M = 66·99$ inches.

It is evident that the form of the relation (7.5) is quite general: if there are r series of observations X_1, X_2, . . . X_r, the mean M of the whole series is related to the means M_1, M_2, . . . M_r of the component series by the equation

$$NM = N_1 M_1 + N_2 M_2 + \ . \ . \ . \ + N_r M_r \qquad . \qquad . \quad (7.6)$$

For the convenient checking of arithmetic, it is useful to note that, if the same arbitrary origin A for the deviations ξ be taken in each case, we must have, denoting the component series by the subscripts 1, 2, . . . r as before,

$$S(f\xi) = S(f_1\xi_1) + S(f_2\xi_2) + \ . \ . \ . \ + S(f_r\xi_r) \qquad . \qquad . \quad (7.7)$$

The agreement of these totals accordingly checks the work.

As an important corollary to the general relation (7.6), it may be noted that the approximate value for the mean obtained from any frequency-distribution is the same whether we assume (1) that all the values in any class are identical with the mid-value of the class-interval, or (2) that the mean of the values in the class is identical with the mid-value of the class-interval.

(c) The mean of all the sums or differences of corresponding observations in two series (of equal numbers of observations) is equal to the sum or difference of the means of the two series.

This follows almost at once. For if

$$X = X_1 \pm X_2$$
$$S(X) = S(X_1) \pm S(X_2)$$

That is, if M, M_1, M_2 be the respective means,

$$M = M_1 \pm M_2 \quad . \qquad . \qquad . \qquad . \quad (7.8)$$

Evidently the form of this result is again quite general, so that if

$$X = X_1 \pm X_2 \pm \ldots \pm X_r$$
$$M = M_1 \pm M_2 \pm \ldots \pm M_r \qquad . \qquad . \qquad . \quad (7.9)$$

As a useful illustration of equation (7.8), consider the case of measurements of any kind that are subject (as indeed all measures must be) to greater or less errors. The actual measurement X in any such case is the algebraic sum of the true measurement X_1 and an error X_2. The mean of the actual measurements M is therefore the sum of the true mean M_1, and the arithmetic mean of the errors M_2. If, and only if, the latter be zero, will the observed mean be identical with the true mean. Errors of grouping (7.14) are a case in point.

The Median.

7.17. The median may be defined as the middlemost or central value of the variable when the values are ranged in order of magnitude, or as the value such that greater and smaller values occur with equal frequency. In the case of a frequency-curve, the median may be defined as that value of the variable the vertical through which divides the area of the curve into two equal parts, as the vertical through Mi in fig. 7.2.

The median, like the mean, fulfils the conditions (b) and (c) of 7.5, seeing that it is based on all the observations made, and that it possesses the simple property of being the central or middlemost value, so that its nature is obvious.

7.18. But the definition does not necessarily lead in all cases to a determinate value. If there be an odd number of different values of X observed, say $2n + 1$, the $(n + 1)$th in order of magnitude is the only value fulfilling the definition. But if there be an even number, say $2n$ different values, any value between the nth and $(n + 1)$th fulfils the conditions. In such a case it appears to be usual to take the mean of the nth and $(n + 1)$th values as the median, but this is a convention supplementary to the definition.

7.19. It should also be noted that in the case of a discontinuous variable the second form of the definition in general breaks down : if we range the values in order there is always a middlemost value (provided the number of observations be odd), but there is not, as a rule, any value such that greater and less values occur with equal frequency. Thus, in Table 6.2 we see that 45 per cent. of the poppy capsules had 12 or fewer stigmatic rays, 55 per cent. had 13 or more ; similarly, 61 per cent. had 13 or fewer rays, 39 per cent. had 14 or more. There is no number of rays such that the frequencies in excess and defect are equal. In the case of the buttercups of Exercise 6.5 (d), page 110, there is no number of petals that even remotely fulfils the required condition. An analogous difficulty may arise, it may be remarked, even in the case of an odd number of observations of a continuous variable if the number of observations be small and several of the observed values identical.

The median is therefore a form of average of most uncertain meaning in cases of strictly discontinuous variation, for it may be exceeded by 5, 10, 15 or 20 per cent. only of the observed values, instead of by 50 per cent. : its use in such cases is to be deprecated, and is perhaps best avoided in any

case, whether the variation be continuous or discontinuous, in which small series of observations have to be dealt with.

Determination of the Median.

7.20. When all the values of the variate are given and the total frequency is small, the median can be determined by inspection as the middlemost value or, if there is no such value, as the mean of the two middlemost values. When the distribution is given as a frequency-distribution, however, a certain amount of approximation is necessary, as in the case of the calculation of the mean.

For the frequency-distribution of a continuous variable a sufficiently approximate value of the median can be obtained by interpolation. If the total frequency is large it is sufficient to assume that the values in each class are uniformly distributed throughout the interval.

Example 7.3.—Let us determine the median of the distribution whose mean we found in Example 7.1. The work may be indicated thus :

Half the total number of observations (8585) . $= 4292 \cdot 5$
Total frequency under $66\frac{15}{16}$ inches . . . $= 3589$

Difference . . . . $= 703 \cdot 5$
Frequency in next interval . . . . $= 1329$

Hence we take the median to be :

$$66\frac{15}{16} + \frac{703 \cdot 5}{1329} \times 1$$
$$= 67 \cdot 47 \text{ inches}$$

The difference between the median and mean in this case is therefore only about one-hundredth of an inch.

Example 7.4.—To find the median of the distribution of Example 7.2.

Half the total number of observations . $= 364 \cdot 5$
Total frequency under 5 years . . . $= 322$

Difference . . . . $= 42 \cdot 5$
Frequency in next interval . . . $= 213$

Hence we take the median to be :

$$5 + \frac{42 \cdot 5}{213} \times 5$$
$$= 6 \text{ years}$$

Here the median is very far from coinciding with the mean.

Graphical Determination of the Median.

7.21. Graphical interpolation may, if desired, be substituted for arithmetical interpolation. Taking the figures of Example 7.1, we see that the number of men with height less than $65\frac{15}{16}$ is 2366, less than $66\frac{15}{16}$ is 8589, less than $67\frac{15}{16}$ is 4918, and less than $68\frac{15}{16}$ is 6148.

Plot the numbers of men with height not exceeding each value of X

to the corresponding value of X on squared paper, to a good large scale, as in fig. 7.3, and draw a smooth curve through the points thus obtained, preferably with the aid of one of the " curves," splines or flexible curves sold by instrument-makers for the purpose.   The point at which the smooth curve so obtained cuts the horizontal line corresponding to a
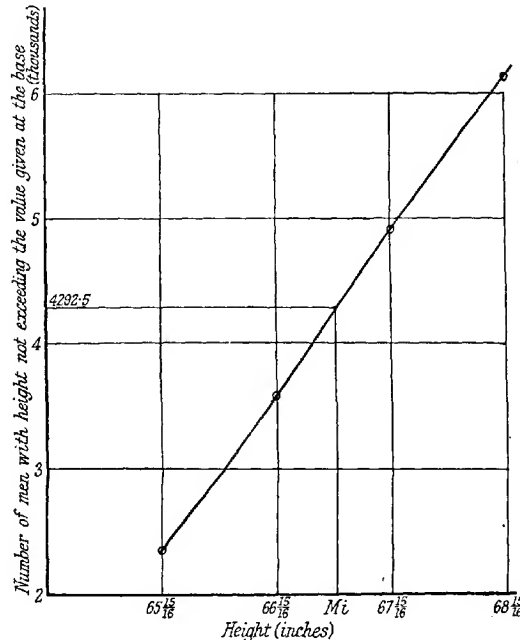
FIG. 7.3.— Determination of the Median by Graphical Interpolation.

total frequency $N/2 = 4292\cdot5$ gives the median.   In general the curve is so flat that the value obtained by this graphical method does not differ appreciably from that calculated arithmetically (the arithmetical process assuming that the curve is a straight line between the points on either side of the median) ; if the curvature is considerable, the graphical value —assuming, of course, careful and accurate draughtsmanship—is to be preferred to the arithmetical value, as it does not involve the crude assumption that the frequency is *uniformly* distributed over the interval in which the median lies.

Comparison of the Mean and the Median.

7.22.   If we adopt the convention that the median of an even number of observations is midway between the two central values, both the

mean and the median satisfy the first three of the desiderata we enumerated in 7.5; that is to say, they are rigidly defined, based on all the observations, and are readily comprehensible. In the remaining three, however, they differ considerably.

7.23. As regards ease of calculation, the median has distinct advantages over the mean.

Whether the stability of the median under fluctuations of sampling is greater than that of the mean depends to some extent on the form of the distribution which is being sampled. In general, the mean is the more stable, but cases occur in which the median is preferable (cf. 7.24 (d) below, and Chap. 20).

When, however, the ease of algebraical treatment of the two forms of average is compared, the superiority lies wholly on the side of the mean. As was shown in 7.16, when several series of observations are combined into a single series, the mean of the resultant distribution can be simply expressed in terms of the means of the components. Expression of the median of the resultant distribution in terms of the medians of the components is, however, not merely complex and difficult, but usually impossible : the value of the resultant median depends on the forms of the component distributions, and not on their medians alone. If two symmetrical distributions of the same form and with the same numbers of observations, but with different medians, be combined, the resultant median must evidently (from symmetry) coincide with the resultant mean, i.e. lie half-way between the means of the components. But if the two components be asymmetrical, or (whatever their form) if the degrees of dispersion or numbers of observations in the two series be different, the resultant median will not coincide with the resultant mean, nor with any other simply assignable value. It is impossible, therefore, to give any theorem for medians analogous to equations (7.5) and (7.6) for means. It is equally impossible to give any theorem analogous to equations (7.8) and (7.9) of 7.16. The median of the sum or difference of pairs of corresponding observations in two series is not, in general, equal to the sum or difference of the medians of the two series ; the median value of a measurement subject to error is not necessarily identical with the true median, even if the median error be zero, i.e. if positive and negative errors be equally frequent.

7.24. These limitations render the applications of the median in any work in which theoretical considerations are necessary comparatively circumscribed. On the other hand, the median may have an advantage over the mean for special reasons.

(a) It is very readily calculated ; a factor to which, however, as already stated, too much weight ought not to be attached.

(b) It is readily obtained, without the necessity of measuring all the objects to be observed, in any case in which the objects can be arranged in order of magnitude. If, for instance, a number of men be ranked in order of stature, the stature of the middlemost is the median, and he alone need be measured. (On the other hand, it is useless in the cases cited at the end of 7.8; the median wage cannot be found from the total of the wages-bill, and the total of the wages-bill is not known when the median is given.)

(c) It is sometimes useful as a makeshift, when the observations are

so given that the calculation of the mean is impossible, owing, *e.g.*, to a final indefinite class.

(*d*) The median *may* sometimes be preferable to the mean, owing to its being less affected by abnormally large or small values of the variable. The stature of a giant would have no more influence on the median stature of a number of men than the stature of any other man whose height is only just greater than the median. If a number of men enjoy incomes closely clustering round a median of £500 a year, the median will be no more affected by the addition to the group of a man with an income of £50,000 than by the addition of a man with an income of £5000, or even £600. If observations of any kind are liable to present occasional *greatly* outlying values of this sort (whether real, or due to errors or blunders), the median will be more stable and less affected by fluctuations of sampling than the arithmetic mean (*cf.* Chap. 20).

(*e*) It may be added that the median is, in a certain sense, a particularly real and natural form of average, for the object or individual that is the median object or individual on any one system of measuring the character with which we are concerned will remain the median on any other method of measurement which leaves the objects in the same relative order. Thus a batch of eggs representing eggs of the median price, when prices are reckoned at so much per dozen, will remain a batch representing the median price when prices are reckoned at so many eggs to the shilling.

The Mode.

7.25. The mode is the value of the variable corresponding to the maximum of the ideal curve which gives the closest possible fit to the actual distribution. It represents the value which is most frequent or typical, the value which is, in fact, the fashion (*la mode*).[1] The mode is sometimes denoted by writing the sign $\smile$ over the variate symbol, *e.g.* $\breve{X}$ means the mode of the values $X_1, X_2, \ldots X_N$.

There is evidently something anticipatory about this definition, for we have not yet defined what we mean by " closest possible fit." For the present the student must content himself with intuitive ideas on this head. Nor have we given a method of finding the curve of closest fit, which would be a necessary preliminary to ascertaining the mode.

7.26. It is, in fact, difficult to determine the mode for such distributions as arise in practice, particularly by elementary methods. It is no use giving merely the mid-value of the class-interval into which the greatest frequency falls, for this is entirely dependent on the choice of the scale of class-intervals. It is no use making the class-intervals very small to avoid error on that account, for the class-frequencies will then become small and the distribution irregular. What we want to arrive at is the mid-value of the interval for which the frequency would be a maximum, if the intervals could be made indefinitely small, and at the same time the number of observations be so increased that the class-

[1] Unless we state expressly to the contrary, we shall be thinking of single-humped distributions in talking of "the" mode. When the distribution is of the complicated form of fig. 6.17 there may be more than one mode. Such distributions are therefore sometimes called multimodal. The mean and the median are still unique for such distributions.

frequencies should run smoothly. As 'the observations cannot, in a practical case, be indefinitely increased, it is evident that some process of smoothing out the irregularities that occur in the actual distribution must be adopted, in order to ascertain the approximate value of the mode. But there is only one smoothing process that is really satisfactory, in so far as every observation can be taken into account in the determination, and that is the method of fitting an ideal frequency-curve of given equation to the actual figures. The value of the variable corresponding to the maximum of the fitted curve is then taken as the mode, in accordance with our definition. The determination of the mode by this—the only strictly satisfactory—method must, however, be left to the more advanced student. The methods of curve-fitting which we shall discuss in Chapter 17 are not appropriate to the fitting of frequency-curves, but we give an approximate method which is of use in certain cases in **24.21.**

Empirical Relation between Mean, Median and Mode.

7.27. For a symmetrical distribution, mean, median and mode coincide, as will be evident on a little consideration. For other distributions, as a rule, they do not. Fig. 7.2 shows the position of the three in a moderately skew distribution.

There is an approximate relation between mean, median and mode which appears to hold good with surprising closeness for moderately asymmetrical distributions, approaching the ideal type of fig. 6.7, and it is one that should be borne in mind as giving—roughly, at all events— the relative values of these three averages for a great many cases with which the student will have to deal. It is expressed by the equation

$$\text{Mode} = \text{Mean} - 3(\text{Mean} - \text{Median})$$

That is to say, the median lies one-third of the distance mean to mode from the mean towards the mode.

The following table gives the true mode and the mode calculated in accordance with the above formula for certain skew distributions of the type of fig. 6.10 : —

Comparison of the Approximate and True Modes in the Case of Five Distributions of the Height of the Barometer for Daily Observations at the Stations named. (Distributions given by Karl Pearson and Alice Lee, *Phil. Trans.,* A, vol. 190, 1897, p. 423.)

Station.	Mean.	Median.	Approximate Mode.	True Mode.
Southampton .	29·981	30·000	30 038	30·039
Londonderry .	29·891	29·915	29·963	29·960
Carmarthen .	2)·952	29·974	30·018	30·013
Glasgow . .	29 886	29·906	29·946	29·967
Dundee . .	29·870	29·890	29·930	29 951

It will be seen that the true and approximate values are extremely close, except in the case of Dundee and Glasgow, where the divergence reaches two-hundredths of an inch.

7.28. Summing up the preceding paragraphs, we may say that the mean is the form of average to use for all general purposes ; it is simply calculated, its value is nearly always determinate, its algebraic treatment is

particularly easy, and in most cases it is rather less affected than the median by errors of sampling. The median is, it is true, somewhat more easily calculated from a given frequency-distribution than is the mean ; it is sometimes a useful makeshift, and in a certain class of cases it is more and not less stable than the mean ; but its use is undesirable in cases of discontinuous variation, its value may be indeterminate, and its algebraic treatment is difficult and often impossible. The mode, finally, is a form of average hardly suitable for elementary use, owing to the difficulty of its determination, but at the same time it represents an important value of the variable. The arithmetic mean should invariably be employed unless there is some very definite reason for the choice of another form of average, and the elementary student will do very well if he limits himself to its use. Objection is sometimes taken to the use of the mean in the case of asymmetrical frequency-distributions, on the ground that the mean is not the mode, and that its value is consequently misleading. But no one in the least degree familiar with the manifold forms taken by frequency-distributions would regard the two as in general identical ; and while the importance of the mode is a good reason for stating its value in addition to that of the mean, it cannot replace the latter. The objection, it may be noted, would apply with almost equal force to the median, for, as we have seen (7.27), the difference between mode and median is usually about two-thirds of the difference between mode and mean.

The Geometric Mean.

7.29. The geometric mean G of a series of values $X_1, X_2, X_3, \ldots X_N$ is defined by the relation

$$G = (X_1 X_2 X_3 \ldots X_N)^{1/N} \qquad . \qquad . \qquad . \quad (7.10)$$

The definition may also be expressed in terms of logarithms :

$$\log G = \frac{1}{N} S(\log X) \qquad . \qquad . \qquad . \quad (7.11)$$

that is to say, the logarithm of the geometric mean of a series of values is the arithmetic mean of their logarithms.

The geometric mean of a given series of quantities is always less than their arithmetic mean ; the student will find a proof in most textbooks of algebra, and in ref. (105). The magnitude of the difference depends largely on the amount of dispersion of the variable in proportion to the magnitude of the mean (*cf.* Exercise 8.12, p. 153). It is necessarily zero, it should be noticed, if even a single value of X is zero, and it may become imaginary if negative values occur.

Calculation of the Geometric Mean.

7.30. From equation (7.11) it will be evident that the calculation of the geometric mean is exactly the same as that of the arithmetic mean, except that instead of adding the values of the variable we add the logarithms of those values. If there are many values we can draw up a frequency table for the logarithms and proceed as in Examples 7.1 and 7.2.

Properties of the Geometric Mean.

7.31. The geometric mean is rigidly defined and takes account of all the observations. It is also fairly easily calculated, though not so easily as the arithmetic mean. It has, however, no simple and obvious properties which render its general nature readily comprehensible. This, coupled with its rather abstract mathematical character, has prevented it from coming into general use as a representative average.

7.32. At the same time, as the following examples show, the geometric mean possesses some important properties, and is readily treated algebraically in certain cases.

(a) If the series of observations X consist of r component series, there being N_1 observations in the first, N_2 in the second, and so on, the geometric mean G of the whole series can be readily expressed in terms of the geometric means G_1, G_2, etc., of the component series. For evidently we have at once (as in **7.16** (b)):

$$N \log G = N_1 \log G_1 + N_2 \log G_2 + \ldots + N_r \log G_r \qquad . \quad (7.12)$$

(b) The geometric mean of the ratios of corresponding observations in two series is equal to the ratio of their geometric means. For if

$$X = X_1/X_2$$
$$\log X = \log X_1 - \log X_2$$

then summing for all pairs of X_1's and X_2's :

$$G = G_1/G_2 \qquad . \qquad . \qquad . \qquad . \quad (7.13)$$

(c) Similarly, if a variable X is given as the product of any number of others, i.e. if

$$X = X_1 X_2 X_3 \ldots X_r$$

X_1, X_2, ... X_r denoting corresponding observations in r different series, the geometric mean G of X is expressed in terms of the geometric means G_1, G_2, ... G_r of X_1, X_2, ... X_r, by the relation

$$G = G_1 G_2 G_3 \ldots G_r \qquad . \qquad . \qquad . \quad (7.14)$$

That is to say, the geometric mean of the product is the product of the geometric means.

7.33. The geometric mean finds applications in several cases where we have to deal with a quantity whose changes tend to be directly proportional to the quantity itself, e.g. populations ; or where we are dealing with an average of ratios, as in index-numbers of prices. Suppose, for instance, we wish to estimate the numbers of a population midway between two epochs (say two census years) at which the population is known. If nothing is known concerning the increase of the population save that the numbers recorded at the first census were P_0 and at the second census n years later P_n, the most reasonable assumption to make is that the percentage increase in each year has been the same, so that the populations in successive years form a geometric series, $P_0 r$ being the population a year after the first census, $P_0 r^2$ two years after the first census, and so on, so that

$$P_n = P_0 r^n \qquad . \qquad . \qquad . \qquad . \quad (7.15)$$

The population midway between the two censuses is therefore

$$P_{n/2} = P_0 r^{n/2} = (P_0 P_n)^{\frac{1}{2}} \qquad . \qquad .. \qquad . \quad (7.16)$$

i.e. the geometric mean of the numbers given by the two censuses. This result must, however, be used with discretion. The rate of increase of population is not necessarily, or even usually, constant over any considerable period of time : if it were so, a curve representing the growth of population as in fig. 7.4 would be everywhere convex to the base, whether
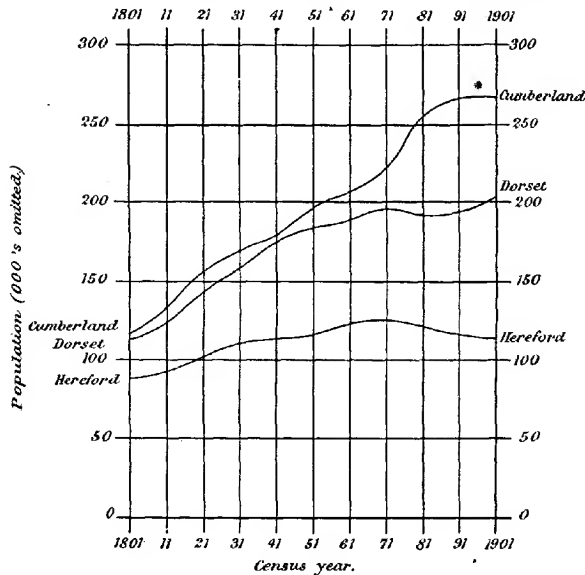
FIG. 7.4.-- Showing the Populations of certain Rural Counties of England for Each Census Year from 1801 to 1901.

the population were increasing or decreasing. In the diagram it will be seen that the curves are frequently concave towards the base, and similar results will often be found for districts in which the population is not increasing very rapidly, and from which there is much emigration. Further, the assumption is not self-consistent in any case in which the rate of increase is not uniform over the entire area—and almost any area can be analysed into parts which are not similar in this respect. For if in one part of the area considered the initial population is P_0 and the common ratio R, and in the remainder of the area the initial population is p_0 and the common ratio r, the population in year n is given by

$$P_n + p_n = P_0 R^n + p_0 r^n$$

This does not represent a constant rate of increase unless $R = r$. If then,

for example, a constant percentage rate of increase be assumed for England and Wales as a whole, it cannot be assumed for the Counties: if it be assumed for the Counties, it cannot be assumed for the country as a whole. The student is referred to refs. (116) and (117) for a discussion of methods that may be used for the consistent estimation of populations in such circumstances.

Use of the Geometric Mean in Index-numbers.

7.34.   The property of the geometric mean illustrated by equation (7.13) renders it, in some respects, a peculiarly convenient form of average in dealing with ratios, *i.e.* " index-numbers," as they are termed, of prices.[1] Let

$$X_0', \quad X_0'', \quad X_0''', \quad \ldots X_0^n$$
$$X_1', \quad X_1'', \quad X_1''', \quad \ldots X_1^n$$
$$X_2', \quad X_2'', \quad X_2''', \quad \ldots X_2^n$$

denote the prices of N commodities in the years $0, 1, 2. \ldots$ Further, let $Y_{10} = X_1/X_0$, and so on, so that

$$Y_{10}', \quad Y_{10}'', \quad Y_{10}''', \quad \ldots Y_{10}^n$$
$$Y_{20}', \quad Y_{20}'', \quad Y_{20}''', \quad \ldots Y_{20}^n$$

represent the ratios of the prices of the several commodities in years *1, 2,* $\ldots$ to their prices in year *0*. These ratios, in practice multiplied by 100, are termed *index-numbers* of the prices of the several commodities, on the year *0* as base.   Evidently some form of average of the Y's for any given year will afford an indication of the general level of prices for that year, provided the commodities chosen are sufficiently numerous and representative.   The question is, what form of average to choose.   If the geometric mean be chosen, and G_{10}, G_{20} denote the geometric means of the Y's for the years *1* and *2* respectively, we have :

$$
\begin{aligned}
\frac{G_{20}}{G_{10}} &= \left(\frac{Y_{20}'}{Y_{10}'} \cdot \frac{Y_{20}''}{Y_{10}''} \cdot \frac{Y_{20}'''}{Y_{10}'''} \cdots \frac{Y_{20}^n}{Y_{10}^n} \right)^{1/N} \\
&= \left(\frac{X_2'}{X_1'} \cdot \frac{X_2''}{X_1''} \cdot \frac{X_2'''}{X_1'''} \cdots \frac{X_2^n}{X_1^n} \right)^{1/N} \\
&= \left(Y_{21}' \cdot Y_{21}'' \cdot Y_{21}''' \cdots Y_{21}^n \right)^{1/N}
\end{aligned}
\qquad . \qquad . \quad (7.17)
$$

From the first form of this equation we see that the ratio of the geometric mean index-number in year *2* to that in year *1* is identical with the geometric mean of the ratios for the index-numbers of the several commodities. A similar property does not hold for any other form of average : the ratio of the arithmetic mean index-numbers is not the same as the arithmetic mean of the ratios, nor is the ratio of the medians the median of the ratios. From the second and third forms of the equation it appears further that the ratio of the geometric mean index-number in year *2* to that in year *1* is independent of the prices in the year first chosen as base (*i.e.* year *0*), and

---

[1] The literature of index-numbers is extensive and it is impossible to discuss them in the limits of this book.   There is still difference of opinion as to the most suitable form of an index-number, and we do not mean to prejudge this question in the above section.

is identical with the geometric mean of the index-numbers for year *2*, on year *1* as base. Again, a similar property does not hold for any other form of average. If arithmetic means of the index-numbers be taken, for example, the ratio of the mean in year *2* to the mean in year *1* will vary with the year taken as base, and will differ more or less from the arithmetic mean ratio of the prices in year *2* to the prices of the same commodities in year *1*; the same statement is true if medians be used. The results given by the use of the geometric mean possess, therefore, a certain consistency that is not exhibited if other forms of average are employed. It was used in a classical paper by Jevons (ref. (108)), though not on quite the same grounds, but has never been at all generally employed, although it is now in use for the index of wholesale prices compiled by the British Board of Trade.

The Harmonic Mean.

7.35. The harmonic mean of a series of quantities is the reciprocal of the arithmetic mean of their reciprocals; that is, if H be the harmonic mean,

$$\frac{1}{H} = \frac{1}{N} S\left(\frac{1}{X}\right). \qquad . \qquad . \qquad . \qquad . \quad (7.18)$$

The following illustration will serve to show the method of calculation:—

Example 7.5.—The table gives the number of litters of mice, in certain breeding experiments, with given numbers (X) in the litter. (Data from A. D. Darbishire, *Biometrika*, vol. 3, pp. 30, 31.)

Number in Litter. X.	Number of Litters. f.	f/X.
1	7	7·000
2	11	5·500
3	16	5·333
4	17	4·250
5	26	5·200
6	31	5·167
7	11	1·571
8	1	0·125
9	1	0·111
—	121	34·257

Whence
$$\frac{1}{H} = \frac{34·257}{121} = 0·2831$$

$$H = 3·532$$

The arithmetic mean is 4·587, more than a unit greater.

Reciprocal Character of Arithmetic and Harmonic Means.

7.36. Prices may be stated in two different ways which are reciprocally related, the resulting arithmetic mean of the one being the harmonic mean of the other. Supposing we had 100 returns of retail prices of eggs, 50 returns showing twelve eggs to the shilling, 30 fourteen to the shilling and 20 ten to the shilling; then the mean number per shilling would be

12·2, equivalent to a price of 0·984d. per egg. But if the prices had been quoted in the form usual for other commodities, we should have had 50 returns showing a price of 1d. per egg, 30 showing a price of 0·857d. and 20 a price of 1·2d. : arithmetic mean 0·997d., a slightly greater value than the harmonic mean of 0·984d.

The harmonic mean of a series of quantities is always lower than the geometric mean of the same quantities, and *a fortiori*, lower than the arithmetic mean, the amount of difference depending largely on the magnitude of the dispersion relatively to the magnitude of the mean (*cf.* Exercise 8.13, p. 153).

SUMMARY.

1. Measures of the location or position of a frequency-distribution are called averages.

2. There are three types of average in general use, the mean (arithmetic, geometric and harmonic), the median and the mode.

3. The arithmetic mean of N values $X_1, X_2, \ldots X_N$ is given by

$$M = \frac{1}{N}S(X)$$

The geometric mean is given by

$$G = (X_1 \ldots X_N)^{1/N}$$

or

$$\log G = \frac{1}{N}S(\log X)$$

The harmonic mean is given by

$$\frac{1}{H} = \frac{1}{N}S\left(\frac{1}{X}\right)$$

4. The median is the central value of the variable when the values are ranged in order of magnitude ; if the number of values is even, the median is conventionally taken to be the arithmetic mean of the two central values.

5. The mode is the value of the variate corresponding to the maximum of the ideal curve which gives the closest possible fit to the actual distribution.

6. For distributions of moderate skewness there is an empirical relationship between the mean, the median and the mode expressed by the equation

$$\text{Mode} = \text{Mean} - 3(\text{Mean} - \text{Median})$$

EXERCISES.

7.1. Verify the following means and medians from the data of Table 6.7, page 94 :—

Stature in Inches for Adult Males in

	England.	Scotland.	Wales.	Ireland.
Mean	. 67·31	68·55	66·62	67·78
Median .	. 67·35	68·48	66·56	67·69

In the calculation of the means use the same arbitrary origin as in Example 7.1 and check your work by the method of 7.16 (*b*).

7.2. The mean of 13 numbers is 10, and the mean of 42 other numbers is 16. Find the mean of the 55 numbers taken together.

7.3. Find the mean weight of adult males in the United Kingdom from the data in the last column of Exercise 6.6, page 111. Find the median weight, and hence find the approximate mode by the relation of 7.27.

7.4. Similarly, find the mean, median and approximate value of the mode for the distribution of fecundity in race-horses, Table 6.9, page 98.

7.5. Using a graphical method, find the median income subject to sur- or super-tax in the financial year 1931 from the data of Table 6.5, page 89.

7.6. Find the arithmetic mean of the first n natural numbers and show that it coincides with the median.

7.7. (Data from *Agricultural Statistics, England and Wales*, Part 2, 1932.) The figures in columns 1 and 2 of the small table below show the index-numbers of prices of certain commodities in the harvest years 1926 and 1931, the years 1911–13 being taken as 100. In column 3 have been added the ratios of the index-numbers in 1931 to those in 1926, the latter being taken as 100.

Find the average ratio of prices in 1931 to those in 1926—

    (1) From the arithmetic mean of the ratios in column 3.
    (2) From the ratio of the arithmetic means of columns 1 and 2.
    (3) From the ratio of the geometric means of columns 1 and 2.
    (4) From the geometric mean of the ratios of column 3.

Note that, by 7.32, the last two methods must give the same result.

Commodity.	Index-number of Price in		Ratios.
	1926.	1931.	31/26.
	1.	2.	3.
1. Wheat. . . .	157	79	50·3
2. Fat Cattle . .	131	118	90·1
3. Milk . . .	163	139	85·3
4. Eggs . . .	149	110	73·8
5. Fruit . . .	165	132	80·0
6. Vegetables . .	135	158	117·0

7.8. Find the arithmetic and geometric means of the series 1, 2, 4, 8, 16, . . . 2^n. Find also the harmonic mean.

7.9. Supposing the frequencies of values 0, 1, 2, . . . of a variable to be given by the terms of the binomial series

$$q^n, \qquad nq^{n-1}p, \qquad \frac{n(n-1)}{1 \cdot 2}q^{n-2}p^2, \quad \ldots$$

where $p + q = 1$, find the mean.

7.10. Show that, in finding the arithmetic mean of a set of readings on a thermometer, it does not matter whether we measure temperature in Centigrade or Fahrenheit degrees, but that in finding the geometric mean it does matter.

7.11. (Data from Census of 1901.) The table below shows the population of the rural sanitary districts of Essex, the urban sanitary districts (other than the borough of West Ham), and the borough of West Ham, at the censuses of 1891 and 1901. Estimate the total population of the county at a date midway between the two censuses, (1) on the assumption that the percentage rate of

increase is constant for the county as a whole; (2) on the assumption that the percentage rate of increase is constant in each group of districts and the borough of West Ham.

Essex.	Population.	
	1891.	1901.
Rural districts . . .	232,867	240,776
West Ham	204,903	267,358
Other urban districts . .	345,604	575,864
Total .	783,374	1,083,998

7.12. (Data from *Agricultural Statistics*, Part 2, 1932.) The following statement shows the monthly average prices of eggs in England and Wales in 1932, as compiled from returns from certain markets for National Mark Specials and English Ordinaries, First Quality, per 120:—

Month.	N.M. Specials.	English Ordinaries, First Quality.
	s. d.	s. d.
January	18 11	15 2
February	15 0	12 11
March	11 11	10 0
April	10 10	9 2
May	10 9	8 9
June	12 0	10 0
July	14 2	12 6
August	15 6	13 9
September . . .	18 10	16 3
October	20 9	18 9
November	24 1	21 8
December	21 2	16 10
Mean for year . .	16 2	13 10

What would have been the mean price for the year in each case if the wholesale prices had been recorded as retail prices sometimes are, *i.e.* at so many eggs per shilling? State your answer in the form of the equivalent price per 120, and obtain it in the shortest way by taking the harmonic mean of the above prices.

CHAPTER 8.

MEASURES OF DISPERSION.

Range.

8.1. We can now turn to a consideration of measures of the dispersion of variate values about the central values we have discussed in the last chapter.

The simplest possible measure of dispersion is the **range**, *i.e.* the difference between the greatest and least values observed. The extreme ease with which this measure may be calculated and its very obvious interpretation have led to its use in many industrial problems. There are, however, serious objections to the use of the range which usually more than offset these advantages.

In the first place, the range is subject to fluctuations of considerable magnitude from sample to sample. There are seldom real upper or lower limits to the values which a variable can take, large or small values being only more or less infrequent. The occurrence of one of these infrequent values may have quite a disproportionate effect on the range. Suppose, for example, we consider the data of Exercise 6.6, page 111, showing the frequency-distributions of weights of adult males in several parts of the United Kingdom. In Wales one individual was observed with a weight of over 280 lb., the next heaviest being under 260 lb. The addition of this one exceptional man to 737 others has increased the range by some 30 lb., or about 20 per cent.

Moreover, the range takes no account of the form of the distribution within the range. We might get the same value for the range from a symmetrical and a J-shaped frequency-curve. Clearly we could not regard two such distributions as exhibiting the same dispersion.

8.2. A measure of dispersion, in fact, should obey conditions similar to those we laid down for measures of location in the last chapter (7.5). That is to say, it should be based on all the observations, should be readily comprehensible, fairly easily calculated, affected as little as possible by fluctuations of sampling, and amenable to algebraical treatment.

There are three measures of dispersion in general use, the **standard deviation,** the **mean deviation** and the **quartile deviation or semi-interquartile range.** We will consider them in that order.

The Standard Deviation.

8.3. The standard deviation is the square root of the arithmetic mean of the squares of all deviations, deviations being measured from the arithmetic mean of the observations. If the standard deviation be denoted by σ, and a deviation from the arithmetic mean by x, then the standard deviation is given by the equation

$$\sigma^2 = \frac{1}{N} S(x^2) . \qquad . \qquad . \qquad . \qquad . \quad (8.1)$$

134

To square all the deviations may seem at first sight an artificial procedure, but it must be remembered that it would be useless to take the mere sum of the deviations, in order to obtain a measure of dispersion, since this sum is necessarily zero if deviations be taken from the mean. In order to obtain some quantity that shall vary with the dispersion, it is necessary to average the deviations by a process that treats them as if they were all of the same sign, and *squaring* is the simplest process for eliminating signs which leads to results of algebraical convenience.

Root-mean-square Deviation.

8.4. The standard deviation is a particular case of a more general quantity, known as the root-mean-square deviation, which has theoretical importance.

Let A be any arbitrary value of X, and let ξ (as in 7.11) denote the deviation of X from A; *i.e.* let

$$\xi = X - A$$

Then we may define the root-mean-square deviation s from the origin A by the equation

$$s^2 = \frac{1}{N} S(\xi^2) \quad . \qquad . \qquad . \qquad . \qquad . \quad (8.2)$$

The standard deviation is the value of the root-mean-square deviation taken from the mean.

8.5. The quantities σ^2 and s^2, *i.e.* the squares of the standard and root-mean-square deviations, are sufficiently important in much theoretical work to have special names.

The square of the standard deviation, σ^2, is called the **variance**.

The quantity $\frac{1}{N} S(\xi^2)$, *i.e.* s^2, is called the **second moment** about the value A. We have already seen (7.11) that the quantity $\frac{1}{N} S(\xi)$ is called the first moment about A, and in the next chapter we shall consider moments of higher orders.

Thus, the variance is the second moment about the mean.

Relation between Standard and Root-mean-square Deviations.

8.6. There is a very simple relation between the standard deviation and the root-mean-square deviation from any other origin. Let

$$M - A = d \quad . \qquad . \qquad . \qquad . \qquad . \quad (8.3)$$

so that

$$\xi = x + d$$

Then

$$\xi^2 = x^2 + 2xd + d^2$$
$$S(\xi^2) = S(x^2) + 2dS(x) + Nd^2$$

But the sum of the deviations from the mean is zero, therefore the second term vanishes, and accordingly

$$s^2 = \sigma^2 + d^2 \quad . \qquad . \qquad . \qquad . \qquad . \quad (8.4)$$

Hence the root-mean-square deviation is least when deviations are measured from the mean, *i.e.* the standard deviation is the least possible root-mean-square deviation.

8.7. If σ and d are the two sides of a right-angled triangle, s is the hypotenuse. If, then, *MH* be the vertical through the mean of a frequency-distribution (fig. 8.1), and *MS* be set off equal to the standard deviation (on the same scale by which the variable X is plotted along the base), *SA* will be the root-mean-square deviation from the point *A*. This construction gives a concrete idea of the way in which the root-mean-square deviation depends on the origin from which deviations are measured. It will be seen that for small values of d the difference of s and σ will be very minute, since A will lie very nearly on the circle drawn through M with centre S and radius *SM* : slight errors in the mean due to approximations in calculation will not, therefore, appreciably affect the value of the standard deviation.
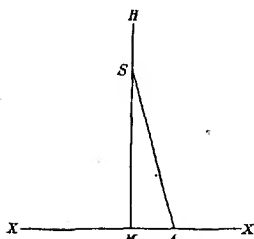
FIG. 8.1.

Calculation of the Standard Deviation.

8.8. If we have to deal with relatively few, say thirty or forty, ungrouped observations, the method of calculating the standard deviation is perfectly straightforward. It is illustrated by the figures below giving the minimum wage-rates for agricultural labourers in England and Wales at the beginning of 1936.

First of all the mean is ascertained. Then we find the values of x by subtracting the mean from all values of the variable. Each difference is squared and the total, $S(x^2)$, obtained. This total divided by the total frequency is the square of the standard deviation.

In practice, we can simplify the arithmetic by working from an arbitrary value A instead of from the mean. Such a value is usually known as the "working mean." When we have found the mean-square deviation s^2 about A we can easily find the value of σ^2 from equation (8.4).

Example 8.1.—Calculation of Standard Deviation for a short series of observations (49) ungrouped. Minimum weekly rates of wages for ordinary adult male agricultural workers in England and Wales as at 1st January 1936.

By inspection of the table opposite we see that the mean is in the neighbourhood of 32 shillings. We therefore take this as the working mean A. The column headed " Difference " is the excess of the value of the variable over this value. The column headed " (Difference)² " is the square of the excess. We find

$$\frac{1}{N}S(\xi) = \frac{-79}{49} = -1.612 \text{ pence}$$

Hence the mean = 32 shillings – 1·612 pence
            = 31 shillings 10·4 pence approximately.

Area.	Wage Rates.		Difference, ξ (pence).	(Difference)^2, ξ^2.
	s.	d.		
Bedford and Huntingdon shires	31	6	− 6	36
Berkshire	31	0	− 12	144
Bucks	32	0	—	—
Cambridgeshire	31	6	− 6	36
Cheshire	32	6	6	36
Cornwall	32	0	—	—
Cumberland	32	6	6	36
Derbyshire	36	0	48	2304
Dorset	31	6	− 6	36
Durham	29	0	− 36	1296
Essex	31	0	− 12	144
Gloucester	31	0	− 12	144
Hampshire	31	0	− 12	144
Hereford	31	0	− 12	144
Hertford	32	0	—	—
Kent	33	0	12	144
Lancashire (South)	32	9	9	81
,, (Rest)	36	6	54	2916
Leicester	33	0	12	144
Lincs (Holland)	34	0	24	576
,, (Kesteven and Lindsey)	31	0	− 12	144
Middlesex	33	8	20	400
Monmouth	32	0	—	—
Norfolk	31	6	− 6	36
Northants	31	6	− 6	36
Northumberland	31	6	− 6	36
Notts	32	0	—	—
Oxfordshire	31	6	− 6	36
Rutland	31	6	6	36
Shropshire	32	0	—	—
Somerset	32	6	6	36
Staffs	31	6	− 6	36
Suffolk	31	0	− 12	144
Surrey	32	3	3	9
Sussex	32	0	—	—
Warwickshire	30	0	− 24	576
Westmorland	31	0	− 12	144
Wiltshire	31	0	− 12	144
Worcester	31	0	− 12	144
Yorks, E. Riding	33	6	18	324
,, N. Riding	33	0	12	144
,, W. Riding	33	9	21	441
Anglesey and Caernarvon	31	0	− 12	144
Carmarthen	31	6	− 6	36
Denbigh and Flint	30	6	− 18	324
Glamôrgan	33	6	18	324
Merioneth and Montgomery	28	6	− 42	1764
Pembroke and Cardigan	31	0	− 12	144
Radnor and Brecon	30	0	− 24	576
Totals	—		− 79	14,539

Also

$$\frac{1}{N}S(\xi^2) = \frac{14,539}{49} = 296{\cdot}714 = s^2$$
$$\sigma^2 = s^2 - d^2 = 296{\cdot}714 - (1{\cdot}612)^2$$
$$= 294{\cdot}112$$
$$\sigma = 17{\cdot}15 \text{ pence approximately.}$$

We would direct the student's attention to the necessity for checking his work at each stage before proceeding to the next. If he neglects this warning he is likely to learn by bitter experience how essential it was. For instance, in the above work it would be well to check the value of the mean by summing the wage rates and dividing by 49. We get in this way:

$$\text{Mean} = \frac{1561s. \ 5d.}{49} = 31s. \ 10{\cdot}4d.$$

which checks with the mean found from the working mean. Secondly, the squares of differences should be checked before they are added, and if the addition is made without a machine, a check should be carried out by summing first from bottom to top and then from top to bottom, to avoid repeating errors. A further systematic check is given in **8.10** below.

8.9. If we have to deal with a grouped frequency-distribution the same artifices and approximations are used as in the calculation of the mean (**7.10** and **7.11**). The mid-value of one of the class-intervals is chosen as the arbitrary origin A from which to measure the deviations ξ, the class-interval is treated as a unit throughout the arithmetic, and all the observations within any one class-interval are treated as if they were identical with the mid-value of the interval. If, as before, we denote the frequency in any one interval by f, these f observations contribute $f\xi^2$ to the sum of the squares of deviations, and we have:

$$s^2 = \frac{1}{N}S(f\xi^2) \ .$$

The standard deviation is then calculated from equation (8.4).

8.10. As the arithmetic in calculating the standard deviation is often extensive, it is as well to use some check similar to that of **7.12**. In this case we have:

$$(\xi + 1)^2 = \xi^2 + 2\xi + 1$$
$$f(\xi + 1)^2 = f\xi^2 + 2f\xi + f$$
$$\therefore \quad S\{f(\xi + 1)^2\} = S(f\xi^2) + 2S(f\xi) + N$$

Hence, if we calculate $S\{f(\xi + 1)^2\}$ as well as $S(f\xi^2)$, the above equation gives us a simple check on the accuracy of our work. The following examples illustrate the method:—

Example 8.2.—Calculation of the Standard Deviation of stature of male adults in the British Isles from the figures of Table 6.7, page 94.

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Height, Inches.	Frequency. f.	Deviation from Value A. ξ.	Product. $f\xi$.	$f(\xi+1)$.	Product. $f\xi^2$.	$f(\xi+1)^2$.
57–	2	– 10	– 20	– 18	200	162
58–	4	– 9	– 36	– 32	324	256
59–	14	– 8	– 112	– 98	896	686
60–	41	– 7	– 287	– 246	2,009	1,476
61–	83	– 6	– 498	– 415	2,988	2,075
62–	169	– 5	– 845	– 676	4,225	2,704
63–	394	– 4	– 1,576	– 1,182	6,304	3,546
64–	669	– 3	– 2,007	– 1,338	6,021	2,676
65–	990	– 2	– 1,980	– 990	3,960	990
66–	1,223	– 1	– 1,223	– 4,995	1,223	—
67–	1,329	0	– 8,584	1,329	—	1,329
68–	1,230	+ 1	1,230	2,460	1,230	4,920
69–	1,063	+ 2	2,126	3,189	4,252	9,567
70–	646	+ 3	1,938	2,584	5,814	10,336
71–	392	+ 4	1,568	1,960	6,272	9,800
72–	202	+ 5	1,010	1,212	5,050	7,272
73–	79	+ 6	474	553	2,844	3,871
74–	32	+ 7	224	256	1,568	2,048
75–	16	+ 8	128	144	1,024	1,296
76–	5	+ 9	45	50	405	500
77–	2	+ 10	20	22	200	242
Total	8,585	—	8,763	13,759	56,809	65,752

$$S(f\xi) = 8,763 - 8,584 = 179$$
$$S\{f(\xi+1)\} = 13,759 - 4,995 = 8,764$$

This is an example we have already considered when calculating the mean, and the work of the first four columns is the same as that of Example 7.1, page 116.

As a check on $S(f\xi)$ we have :

$$S\{f(\xi+1)\} - S(f\xi) = 8764 - 179$$
$$= 8585$$
$$= N$$

As a check on $S(f\xi^2)$ we have :

$$S\{f(\xi+1)^2\} - S(f\xi^2) - 2S(f\xi) = 65,752 - 56,809 - 358$$
$$= 8585$$
$$= N$$

From previous work, $M - A = d = +0.0209$ class-intervals or inches.

$$\frac{S(f\xi^2)}{N} = \frac{56,809}{8585} = 6.6172$$
$$\sigma^2 = 6.6172 - (0.0209)^2$$
$$= 6.6168$$
$$\therefore \quad \sigma = 2.57 \text{ class-intervals or inches.}$$

Example 8.3.—Let us find the mean and standard deviation of the distribution of Australian marriages given in Table 6.8, page 96.

Calculation of Standard Deviation of age of bridegroom in a distribution of Australian marriages.

Age of Bridegroom. (Central Value, Years.)	Frequency. f.	ξ.	$f\xi$.	$f(\xi+1)$.	$f\xi^2$.	$f(\xi+1)^2$.
16·5	294	-4	- 1,176	- 882	4,704	2,646
19·5	10,995	-3	- 32,985	-21,990	98,955	43,980
22·5	61,001	-2	- 122,002	-61,001	244,004	61,001
25·5	73,054	-1	- 73,054	—	73,054	—
28·5	56,501	0	—	56,501	—	56,501
31·5	33,478	1	33,478	66,956	33,478	133,912
34·5	20,569	2	41,138	61,707	82,276	185,121
37·5	14,281	3	42,843	57,124	128,529	228,496
40·5	9,320	4	37,280	46,600	149,120	233,000
43·5	6,236	5	31,180	37,416	155,900	224,496
46·5	4,770	6	28,620	33,390	171,720	233,730
49·5	3,620	7	25,340	28,960	177,380	231,680
52·5	2,190	8	17,520	19,710	140,160	177,390
55·5	1,655	9	14,895	16,550	134,055	165,500
58·5	1,100	10	11,000	12,100	110,000	133,100
61·5	810	11	8,910	9,720	98,010	116,640
64·5	649	12	7,788	8,437	93,456	109,681
67·5	487	13	6,331	6,818	82,303	95,452
70·5	326	14	4,564	4,890	63,896	73,350
73·5	211	15	3,165	3,376	47,475	54,016
76·5	119	16	1,904	2,023	30,464	34,391
79·5	73	17	1,241	1,314	21,097	23,652
82·5	27	18	486	513	8,748	9,747
85·5	14	19	266	280	5,054	5,600
88·5	5	20	100	105	2,000	2,205
Total	301,785	—	88,832	390,617	2,155,838	2,635,287

We take a working mean $A = 28·5$.

As a check on $S(f\xi)$ we have :

$$S\{f(\xi+1)\} - S(f\xi) = 390,617 - 88,832$$
$$= 301,785$$
$$= N$$

As a check on $S(f\xi^2)$ we have :

$$S\{f(\xi+1)^2\} - S(f\xi^2) - 2S(f\xi) = 2,635,287 - 2,155,838 - 177,664$$
$$= 301,785$$
$$= N$$

Then

$$M - A = d = \frac{88,832}{301,785} = 0·29436 \text{ interval}$$
$$= 0·88308 \text{ year}$$

Hence,

$$M = 29·383 \text{ years}$$

We have :

$$s^2 = \frac{2,155,838}{301,785} = 7 \cdot 143622 \text{ intervals}^2$$

$$\sigma^2 = s^2 - d^2 \text{ intervals}^2$$
$$= 7 \cdot 056974 \text{ intervals}^2$$

$$\sigma = 2 \cdot 6565 \text{ intervals}$$
$$= 7 \cdot 969, \text{ or 8 years approximately.}$$

Sheppard's Correction for Grouping.

8.11. The student must remember that the treatment of all the values of a variable in a class-interval as if they were concentrated at the centre of that interval is an approximation, although, for distributions of symmetrical or moderately skew type and class-intervals not greater than about one-twentieth of the range, the approximation may be a very close one.

It has been shown that if

(a) the distribution of frequency is continuous, and
(b) the frequency tapers off to zero in both directions,

the variance obtained from grouped data may with advantage be corrected for the grouping effect by subtracting from it one-twelfth of the square of the class-interval; *i.e.* if the class-interval be h units in width, σ^2 the corrected value of the variance and σ_1^2 the value obtained from the grouped data :

$$\sigma^2 = \sigma_1^2 - \frac{h^2}{12} . \qquad . \qquad . \qquad . \qquad (8.5)$$

The proof of this formula lies outside the scope of this book. We may emphasise condition (b). The Sheppard correction is not applicable to J- or U-shaped distributions, or even to the skew form of fig. 6.7 (b), page 95.

Furthermore, unless the total frequency is fairly large, the Sheppard correction is likely to be of secondary importance compared with fluctuations of sampling (see **21.13**). We suggest that, as a general rule, the correction should not be made unless the frequency is at least 1000, or the grouping coarser than that given by intervals of about one-twentieth of the range. We give in Exercise 8.15 a result which will convey the general magnitude of the correction for the finer grouping.

Example 8.4.—In Example 8.2 we have :

$$\sigma_1^2 = 6 \cdot 6168$$
$$\frac{h^2}{12} = 0 \cdot 0833$$
$$\therefore \text{ Corrected value } \sigma^2 = 6 \cdot 5335$$

and σ corrected $= 2 \cdot 56$, differing from the uncorrected value by $0 \cdot 01$.

Example 8.5.—In Example 8.3 we have :

$$\sigma^2 \text{ (uncorrected)} = 7 \cdot 056974 \text{ intervals}^2$$

Here σ^2 is expressed in terms of h^2, and hence to correct it we subtract $\frac{1}{12}$, giving

$$\sigma^2 \text{ (corrected)} = 6 \cdot 973641$$
$$\sigma = 2 \cdot 6408 \text{ intervals}$$
$$= 7 \cdot 922 \text{ years}$$

as against an uncorrected value of $7 \cdot 969$ years.

Spread of Observations and Standard Deviation.

8.12. It is a useful empirical rule to remember that a range of six times the standard deviation usually includes 99 per cent. or more of all the observations in the case of distributions of the symmetrical or moderately asymmetrical type. Thus in Example 8.2 the standard deviation is $2 \cdot 57$ in., six times this is $15 \cdot 42$ in., and a range from, say, 60 in. to $75 \cdot 4$ in. includes all but some 36 out of 8585 individuals, *i.e.* about $99 \cdot 6$ per cent. This rough rule serves to give a more definite and concrete meaning to the standard deviation, and also to check arithmetical work to some extent—sufficiently, that is to say, to guard against very gross blunders. It must not be expected to hold for short series of observations : in Example 8.1, for instance, the actual range is a good deal less than six times the standard deviation.

Properties of the Standard Deviation.

8.13. The standard deviation is the measure of dispersion which it is most easy to treat by algebraical methods, resembling in this respect the arithmetic mean amongst measures of position. The majority of illustrations of its treatment must be postponed to a later stage (Chap. 16), but the work of 8.6 has already served as one example. We showed in **7.16** that if a series of observations of which the mean is M consists of two component series, of which the means are M_1 and M_2 respectively,

$$NM = N_1 M_1 + N_2 M_2$$

N_1 and N_2 being the numbers of observations in the two component series, and $N = N_1 + N_2$ the number in the entire series. Similarly, the standard deviation σ of the whole series may be expressed in terms of the standard deviations σ_1 and σ_2 of the components and their respective means. Let

$$M_1 - M = d_1$$
$$M_2 - M = d_2$$

Then the mean-square deviations of the component series about the mean M are, by equation (8.4), $\sigma_1^2 + d_1^2$ and $\sigma_2^2 + d_2^2$ respectively. Therefore, for the whole series,

$$N\sigma^2 = N_1(\sigma_1^2 + d_1^2) + N_2(\sigma_2^2 + d_2^2) \qquad . \qquad . \quad (8.6)$$

If the numbers of observations in the component series be equal and the means be coincident, we have as a special case :

$$\sigma^2 = \tfrac{1}{2}(\sigma_1^2 + \sigma_2^2) \qquad . \qquad . \qquad . \qquad . \quad (8.7)$$

so that in this case the square of the standard deviation of the whole

series is the arithmetic mean of the squares of the standard deviations of its components.

It is evident that the form of the relation (8.6) is quite general : if a series of observations consists of r component series with standard deviations σ_1, σ_2, . . . σ_r, and means diverging from the general mean of the whole series by d_1, d_2, . . . d_r, the standard deviation σ of the whole series is given (using m to denote any subscript) by the equation

$$N\sigma^2 = S(N_m\sigma_m{}^2) + S(N_m d_m{}^2) \qquad . \qquad . \qquad . \quad (8.8)$$

Again, as in **7.16**, it is convenient to note, for the checking of arithmetic, that if the same arbitrary origin be used for the calculation of the standard deviations in a number of component distributions, we must have :

$$S(f\xi^2) = S(f_1\xi_1{}^2) + S(f_2\xi_2{}^2) + \ . \ . \ . \ + S(f_r\xi_r{}^2) \quad . \qquad . \quad (8.9)$$

8.14. As another useful illustration, let us find the standard deviation of the first N natural numbers. The mean in this case is evidently $(N+1)/2$. Further, as is shown in any elementary algebra, the sum of the squares of the first N natural numbers is

$$\frac{N(N+1)(2N+1)}{6}$$

Applying equation (8.4) we have that the standard deviation σ is given by

$$\sigma^2 = \tfrac{1}{6}(N+1)(2N+1) - \tfrac{1}{4}(N+1)^2$$

that is,

$$\sigma^2 = \tfrac{1}{12}(N^2 - 1) \qquad . \qquad . \qquad . \qquad . \quad (8.10)$$

This result is of service if the relative merit of, or the relative intensity of some character in, the different individuals of a series is recorded not by means of measurements, *e.g.* marks awarded on some system of examination, but merely by means of the respective positions when ranked in order as regards the character, in the same way as boys are numbered in a class. With N individuals there are always N *ranks*, as they are termed, whatever the character, and the standard deviation is therefore always that given by equation (8.10).

Another useful result follows at once from equation (8.10), namely, the standard deviation of a frequency-distribution in which all values of X within a range $\pm l/2$ on either side of the mean are equally frequent, values outside these limits not occurring, so that the frequency-distribution may be represented by a rectangle. The base l may be supposed divided into a very large number N of equal elements, and the standard deviation reduces to that of the first N natural numbers when N is made indefinitely large. The single unit then becomes negligible compared with N, and consequently

$$\sigma^2 = \frac{l^2}{12} \quad . \qquad . \qquad . \qquad . \qquad . \quad (8.11)$$

8.15. It will be seen from the preceding paragraphs that the standard deviation possesses the majority at least of the properties which are desirable in a measure of dispersion as in an average (**7.5**). It is rigidly defined ; it is based on all the observations made ; it is calculated with reasonable ease ; it lends itself readily to algebraical treatment ; and we

may add, though the student will have to take the statement on trust for the present, that it is, as a rule, the measure least affected by fluctuations of sampling. On the other hand, it may be said that its general nature is not very readily comprehended, and that the process of squaring deviations and then taking the square root of the mean seems a little involved. The student will, however, soon surmount this feeling after a little practice in the calculation and use of the constant, and will realise, as he advances further, the advantages that it possesses. Such root-mean-square quantities, it may be added, frequently occur in other branches of science. The standard deviation should always be used as the measure of dispersion, unless there is some very definite reason for preferring another measure, just as the arithmetic mean should be used as the measure of position.

Note on Nomenclature.

8.16. A great deal of confusion has been introduced into statistical literature by the many different expressions which have been used for the standard deviation and simple derivatives of it. It used to be almost a case of *tot homines quot nomina*, and as the student may meet these expressions elsewhere, we give a short list of them. The term " standard deviation " is now almost universally accepted, and in this book we shall use no other.

" Mean error " (Gauss), " mean square error " and " error of mean square " (Airy) have all been used to denote the standard deviation.

The standard deviation is not to be confused with the "standard error." We shall use this term in a special sense, that of the standard deviation of simple sampling (*cf.* **19.8**).

The standard deviation multiplied by the square root of 2 is also known as " the modulus." The student will see the reason for this multiplication later. The reciprocal of the modulus is called the " precision."

There is also a quantity known as the " probable error," which is defined as being 0·67449 times the standard deviation (*cf.* **19.9**). These last four quantities are particularly important in the theory of errors of observation and the theory of sampling.

Finally, we may remark that since ·we shall use the expression " standard deviation " very frequently, we shall sometimes use the abbreviation " s.d." or simply the symbol σ.

Mean Deviation.

8.17. We have already remarked that it would be useless to take the sum of deviations from the mean as a measure of dispersion because such sum is identically zero. We therefore removed the signs of the deviations by squaring to reach the standard deviation.

It is also possible to overcome this difficulty by adding the sum of deviations taken regardless of sign. The arithmetic mean of these " absolute " deviations is called the **mean deviation.**

If we write $|\xi|$ to denote the deviation from an arbitrary value A taken ·as positive whatever its actual sign, the mean deviation is thus defined as

$$\text{m.d.} = \frac{1}{N}S(|\xi|) \qquad . \qquad . \qquad . \qquad . \quad (8.12)$$

(The expression $|\xi|$ is read "mod ξ"—an abbreviation for "the modulus of ξ").

8.18. Just as the root-mean-square deviation is least when deviations are measured from the arithmetic mean, so the mean deviation is least when deviations are measured from the median. For suppose that, for some origin exceeded by m values out of N, the mean deviation has a value Δ. Let the origin be displaced by an amount c until it is just exceeded by $m-1$ of the values only, *i.e.* until it coincides with the mth value from the upper end of the series. By this displacement of the origin the sum of deviations in excess of the origin is reduced by mc, while the sum of deviations in defect of the mean is increased by $(N-m)c$. The new mean deviation is therefore

$$\Delta + \frac{(N-m)c - mc}{N}$$

$$= \Delta + \frac{1}{N}(N - 2m)c$$

The new mean deviation is accordingly less than the old so long as

$$m > \tfrac{1}{2}N$$

That is to say, if N be even, the mean deviation is constant for all origins within the range between the $N/2$th and the $(N/2+1)$th observations; and this value is the least; if N be odd, the mean deviation is lowest when the origin coincides with the $(N+1)/2$th observation. The mean deviation is therefore a minimum when deviations are measured from the median or, if the latter be indeterminate, from an origin within the range in which it lies.

Calculation of the Mean Deviation.

8.19. The mean deviation is perhaps most easily calculated about the mean, which is always determinate, except in the case of distributions with an indeterminate final class. As, however, it is a minimum about the median, we sometimes require to know the value about that point. The following examples will make the method of calculation clear.

Example 8.6.—Let us find the mean deviation about the mean and about the median in the ungrouped data of Example 8.1.

The data were arranged in alphabetical order of the county wage areas, which makes it a little difficult to ascertain the median by inspection. On rearranging in order of magnitude, we find that the median is the value 31s. 6d.

The deviations from the median value are, then, in order of magnitude

-36, -30, -18, -18, -12, -6 (12 times), 0 (10 times), 6 (7 times), 9, 12, 12, 12, 15, 18, 18, 18, 24, 24, 26, 27, 30, 54, 60

The sum of the negative deviations $= -186$
The sum of the positive deviations $=\ \ 401$
Hence the sum of absolute deviations $=\ \ 587$

Hence m.d. $= \dfrac{587}{49} = 12$ pence approximately.

10

To find the m.d. about the mean, 31s. 10·4d., we note that the 27 negative or zero deviations from the median would be increased by 4·4 pence on transferring to the mean, and the 22 positive deviations decreased by 4·4 pence.   The net effect on the total absolute deviations is then an increase of $(27 - 22) \times 4\cdot4$ pence $= 22$ pence.

Hence the m.d. about the mean is :

$$\frac{587}{49} + \frac{22}{49}$$

$$= 12\cdot43 \text{ pence}$$

Example 8.7.—Let us find the mean deviation of heights about the mean in the data of Example 8.2.

In the case of a grouped frequency-distribution the sum of deviations should first be calculated from the centre of the class-interval in which the mean (or median) lies and then reduced to the mean (or median) as origin.

In this case the mean lies in the interval 67–.   We found when calculating it that the negative deviations totalled – 8584 and the positive deviations 8763.   Hence the sum of absolute deviations from the centre of the interval is 17,347—the unit of measurement being the class-interval.

To reduce to the mean as origin we note that if the number of observations below the mean is N_1 and above the mean N_2, and $M - A = d$ as before, we have to add $N_1 d$ to the sum when found and subtract $N_2 d$.   In this case $d = 0\cdot02$ class-interval, $N_1 = 4918$ and $N_2 = 3667$.

Hence, we must add

$$(4918 - 3667) \times 0\cdot02 = + 25 \text{ intervals}$$

i.e.   The total of deviations $= 17,372$

and

$$\text{m.d.} = \frac{17,372}{8,585} = 2\cdot02 \text{ intervals or inches.}$$

The mean deviation from the median should be found in a similar way, the calculation being assisted if the class-interval in which the median lies is taken as origin.

8.20.   As in the case of the standard deviation, the above calculations assume for certain purposes that all the values of the variable can be treated as if they were concentrated at the centres of class-intervals.   This gives sufficient accuracy for all practical purposes if the class-intervals are reasonably narrow.   It has not been found possible to give any simple correction, such as Sheppard's correction, for errors of grouping in the mean deviation, but we give at the end of this chapter an exercise (8.11) as to the correction to be applied if the values in each interval are treated as if they were evenly distributed over the interval instead of being concentrated at its centre.

Empirical Relation between Mean and Standard Deviations for Symmetrical or Moderately Skew Distributions.

* **8.21.**   It is a useful rule for the student to remember that for symmetrical or moderately skew distributions the mean deviation is about

four-fifths of the standard deviation. Thus, for the distribution of male statures of Examples 8.2 and 8.7, we have :

$$\frac{\text{m.d.}}{\text{s.d.}} = \frac{2\cdot02}{2\cdot57} = 0\cdot79$$

For the short series of observations of Example 8.1 :

$$\frac{\text{m.d.}}{\text{s.d.}} = \frac{12\cdot43}{17\cdot15} = 0\cdot72$$

Quartiles.

8.22. A natural extension of the idea of the median consists in ascertaining the variate values Q_1 and Q_3, such that one-quarter of the observations lies below Q_1 and one-quarter above Q_3. In this case clearly one-quarter lies between Q_1 and Mi, the median, and one-quarter between Mi and Q_3.

Q_1 is termed the **lower quartile** and Q_3 the **upper quartile**. The quartiles and the median thus divide the observed values of the variable into four classes of equal frequency.

We saw that if the number of observations was even, there was an indeterminacy in the position of the median which required the additional convention that in such cases the median would be taken to be mid-way between the two central values. Similar indeterminacies may arise in fixing the quartiles unless the number of observations is one less than a multiple of four. Such cases are treated in an analogous way by supplementary conventions, which will be clear from the following examples.

Example 8.8.—To determine the quartiles of the data of Example 8.1.

Here there are 49 observations, and so the 25th gives the median. We regard half the 25th observation as falling below the median and half above. The lower quartile must divide into two equal parts the $24\frac{1}{2}$ observations falling below the median. The observations other than the median are:

28/6, 29/-, 30/-, 30/-, 30/6, 31/- (12 times), 31/6 (7 times).

The lower quartile must divide the $24\frac{1}{2}$ observations into two sets of $12\frac{1}{4}$. The 12th and the 13th values are both, as it happens, 31/-, and Q_1 being between the two is thus 31/- also.

The 24 observations between the median and the highest value are:

31/6 (twice), 32/- (7 times), 32/3, 32/6 (3 times), 32/9, 33/- (3 times), 33/6, 33/6, 33/8, 33/9, 34/-, 36/-, 36/6.

The 12th and 13th observations are both 32/6, and hence this is the value of Q_3.

If the 12th and 13th observations had been, say, 32/6 and 33/-, we might have taken Q_3 to be 32/6 but regarded $\frac{1}{4}$ of the 12th observation as lying above that value.

Example 8.9.—To determine the quartiles of the distribution of Example 8.2.

Data of this kind are treated by simple arithmetical interpolation or graphical interpolation on the lines of **7.20** or **7.21**.

The quartiles are to divide the distribution into four equal parts. We have, therefore,

$$\frac{8585}{4} = 2146 \cdot 25$$

To the interval 65– are   1376   individuals

Difference =   770·25

Hence, Q_1 is $\dfrac{770 \cdot 25}{990}$ inches from the beginning of the interval, which is $64\frac{15}{16}$.

$$\therefore \quad Q_1 = 65 \cdot 71$$

Similarly, from the interval 70– onwards are 1374 individuals.
Difference from $2146 \cdot 25 = 772 \cdot 25$.
Hence,

$$Q_3 = 69\tfrac{15}{16} - \frac{772 \cdot 25}{1063}$$
$$= 69 \cdot 21 \text{ inches}$$

It is left to the student to check the values by graphical interpolation.

Quartile Deviation.

8.23.   If Mi be the value of the median, in a symmetrical distribution

$$Mi - Q_1 = Q_3 - Mi$$

and the difference may be taken as a measure of dispersion.   But as no distribution is rigidly symmetrical, it is usual to take as the measure

$$Q = \frac{Q_3 - Q_1}{2}$$

and Q is termed the quartile deviation, or better, the semi-interquartile range—it is not a measure of the deviation from any particular average.
Thus, from the values calculated in Example 8.8 we have :

$$Q = \frac{32/6 - 31/\text{-}}{2} = \frac{18^d}{2} = 9 \text{ pence}$$

and from Example 8.9 we have :

$$Q = \frac{69 \cdot 21 - 65 \cdot 71}{2} = 1 \cdot 75 \text{ inches}$$

Empirical Relation between Quartile and Standard Deviations.

8.24.   For symmetrical and moderately skew distributions the semi-interquartile range is usually about two-thirds of the standard deviation.
Thus, for the height distribution of Examples 8.2 and 8.9,

$$\frac{Q}{\sigma} = \frac{1 \cdot 75}{2 \cdot 57} = 0 \cdot 68$$

For the wage statistics of Examples 8.1 and 8.8,

$$\frac{Q}{\sigma} = \frac{9}{17 \cdot 15} = 0 \cdot 52$$

which is considerably lower. We should, however, hardly have expected the comparatively few observations comprised in these data to conform at all closely to the empirical relation.

8.25. It follows from this relation that a range of 6 times the standard deviation corresponds to a range of 9 times the semi-interquartile range (and 7·5 times the mean deviation). Within these ranges we expect to find at least 99 per cent. of the observations in symmetrical or moderately skew distributions.

Comparison of the Three Measures of Dispersion.

8.26. The semi-interquartile range has two advantages over the standard deviation and the mean deviation ; it is calculated with great ease, and it has a clear and simple meaning.

In almost all other respects the advantage lies with the standard deviation. The semi-interquartile range has no simple algebraical properties, and its behaviour under fluctuations of sampling is difficult to decide. In all but the most elementary statistical work these are overwhelming disadvantages, and the use of the semi-interquartile range is not to be recommended unless the calculation of the standard deviation has been rendered difficult or impossible, *e.g.* owing to the employment of irregular class-frequencies or of an indefinite terminal class.

Absolute Measures of Dispersion.

8.27. The three measures of dispersion we have been discussing have all been expressed in terms of the units of the variate ; *e.g.* the standard deviation of height-frequencies was found in inches, and the mean deviation of wage-frequencies in pence. It is thus impossible to compare dispersions in different universes unless they happen to be measured in the same units.

For this reason some statisticians have recommended the use of " absolute " measures of dispersion, which shall be pure numbers and not expressible in some particular scale of units. Such measures would permit of comparison between universes of very different natures.

It is easy to construct several coefficients of the kind required. The standard deviation and the mean deviation have the dimensions of a length, and it is only necessary to divide them by another factor which has the same dimensions ; *e.g.*

$$\frac{\text{Mean deviation}}{\text{Mean}}, \quad \frac{\text{Mean deviation}}{\text{Mode}} \quad \text{and} \quad \frac{\text{Standard deviation}}{\text{Mean}}$$

are all of the required type.

Coefficient of Variation.

8.28. The last-mentioned in the foregoing paragraph in a modified form is the only coefficient which has come into general use. We define the **Coefficient of Variation**, v, as

$$v = 100\frac{\sigma}{M} \qquad . \qquad . \qquad . \qquad . \qquad (8.13)$$

This coefficient has been used by Karl Pearson in comparing the relative variations of corresponding organs or characters in the two sexes, and more

recently by G. S. Wilson in researches on the bacteriological grading of milk (ref. (159)).

Reduction of Frequency-distribution to Absolute Scale.

8.29. Comparability of form may, however, be reached in a different way; that is to say, by regarding σ itself as a unit and expressing other measures in terms of it. Thus, in the height distribution of Example 8.2, $\sigma = 2\cdot57$ inches, or 1 inch $= 0\cdot389$ σ. Hence the intervals are $0\cdot389$ σ in width, and run: $57 \times 0\cdot389$ $\sigma -$, $58 \times 0\cdot389$ $\sigma -$, etc.; *i.e.* $22\cdot173$ $\sigma -$, $22\cdot562$ $\sigma -$, etc.

A distribution expressed in this way has unit standard deviation, for

$$\frac{1}{N}S\left(\frac{x}{\sigma}\right)^2 = \frac{1}{\sigma^2 N}S(x^2) = \frac{\sigma^2}{\sigma^2}$$
$$= 1$$

The distribution reduced to the scale of σ may thus be regarded as expressed in "absolute" units, and two distributions expressed in this way may readily be compared as regards form, but not as regards dispersion, for this has been made the same in the two cases.

Deciles and Percentiles.

8.30. We may conclude this chapter by describing briefly methods which have been much used in the past in lieu of the methods described in this and the preceding chapter.

Instead of dividing the total frequency into 4 parts by quartiles, we may divide it into 100 parts by what are called **percentiles**. Or we may divide into 10 parts by **deciles**. The theory of these quantities is precisely analogous to that of the quartiles: there may, for instance, be certain indeterminacies in their exact definition which are removed by supplementary conventions; they can be obtained by arithmetical or graphical interpolation; and they have simple and obvious meanings.

Quantities such as quartiles, deciles, etc., which divide the total frequency into a number of parts, are called *grades*, and when we speak of the grade of an individual we mean thereby the proportion of the total frequency which lies below it. Conventionally, half the individual is regarded as lying above, and half below, the point determined by the variate value which it bears.

8.31. The values of the percentiles may be used to draw what is known as Galton's ogive curve. In fig. 8.2 we have plotted the 100 grades along the horizontal against the height corresponding to any given percentile up the vertical, for the height distribution of Example 8.2. The curve shows what percentage of the universe falls below any specified height.

8.32. An extension of the method to the treatment of non-measurable characters has also become of some importance. For example, the capacity of the different boys in a class as regards some school subject cannot be directly measured, but it may not be very difficult for the master to arrange them in order of merit as regards this character: if the boys are then "numbered up" in order, the number of each boy, or his **rank**, serves as some sort of index to his capacity (*cf.* the remarks in **8.14**). It should be noted that rank in this sense is not quite the same as

grade ; if a boy is tenth, say, from the bottom in a class of a hundred his grade is 9·5, but the method is in principle the same as that of grades or percentiles. The method of ranks, grades or percentiles in such a case may be a very serviceable auxiliary, though, of course, it is better if possible to obtain a numerical measure. But if, in the case of a measurable character, the percentiles are used not merely as constants illustrative of
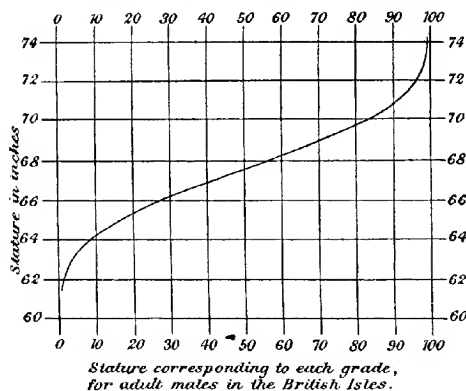
FIG. 8.2.—Ogive Curve for Stature (same data as fig. 6.6, p. 95).

certain aspects of the frequency-distribution, but entirely to replace the table giving the frequency-distribution, serious inconvenience may be caused, as the application of other methods to the data is barred. Given the table showing the frequency-distribution, the reader can calculate not only the percentiles, but any form of average or measure of dispersion that has yet been proposed, to a sufficiently high degree of approximation. But given only the percentiles, or at least so few of them as the nine deciles, he cannot pass back to the frequency-distribution, and thence to other constants, with any degree of accuracy. In all cases of published work, therefore, the figures of the frequency-distribution should be given ; they are absolutely fundamental.

SUMMARY.

1. The standard deviation σ is defined by

$$\sigma^2 = \frac{1}{N}S(x^2)$$

where x is the deviation from the arithmetic mean. σ^2 is called the " variance."

2. The root-mean-square deviation s about a point A is defined by

$$s^2 = \frac{1}{N}S(\xi^2)$$

where ξ is the deviation from A.

3. If $M - A = d$, then

$$s^2 = \sigma^2 + d^2.$$

4. For grouped data the variance should be corrected by subtracting $\frac{h^2}{12}$, where h is the width of the class-interval, provided that (a) the frequency is continuous, and (b) that it tapers off to zero in both directions.

5. The s.d. is the minimum root-mean-square deviation.

6. The mean deviation is defined as

$$\text{m.d.} = \frac{1}{N}S(|\xi|).$$

7. The m.d. is a minimum about the median.

8. The quartiles are the values of the variate which divide the total frequency into 4 equal parts ; similarly, the deciles divide it into 10 equal parts and the percentiles into 100 equal parts.

9. The quartile deviation, or semi-interquartile range, is defined as

$$Q = \frac{Q_3 - Q_1}{2}$$

10. For symmetrical or moderately skew distributions,

$$\text{m.d.} = 0.8\sigma \quad \text{and} \quad Q = 0.67\sigma \text{ approximately.}$$

11. For the majority of such distributions 99 per cent. of the total frequency lies within a range of 6σ, 7.5 m.d. or $9Q$.

EXERCISES.

8.1. Verify the following for the data of Table 6.7, page 94 (in continuation of the work of Exercise 7.1) : –

	Stature in Inches for Adult Males born in			
	England.	Scotland.	Wales.	Ireland.
Standard deviation (uncorrected) .	2·56	2·50	2·35	2·17
Mean deviation	2·05	1·95	1·82	1·69
Quartile deviation	1·78	1·56	1·46	1·35
Mean deviation/standard deviation .	0·80	0·78	0·78	0·78
Quartile deviation/standard deviation.	0·69	0·62	0·62	0·62
Lower quartile.	65·55	66·92	65·06	66·39
Upper ,, ...	69·10	70·04	67·98	69·10

8.2. Find the standard deviation, mean deviation, quartiles and semi-interquartile range for the data in the last column of the table of Exercise 6.6, page 111 (in continuation of the work of Exercise 7.3).

Compare the ratios of mean and quartile deviations to the standard deviation with those stated in 8.21 and 8.24 to be usual for moderately skew distributions.

8.3. Using, or extending if necessary, your diagram for Exercise 7.5, page 132, find the median and upper quartile for incomes subject to sur- or super-tax.

Find also the 9th decile (the value exceeded by 10 per cent. of incomes only).

8.4. Find the quartiles of the distribution of Australian marriages given in Example 8.3, and find the semi-interquartile range.

8.5. Find directly the standard deviation of the natural numbers from 1 to 10, and hence verify equation (8.10).

8.6. Show that, for any distribution, the standard deviation is not less than the mean deviation about the mean.

8.7. Show that, for a ⨼-shaped distribution with the maximum frequency towards the lower values of the variate, the median is nearer to Q_1 than to Q_3.

8.8. Find the mean and standard deviation of the following numbers (1) without further grouping, (2) grouping the numbers by fives (40–, 45–, 50–, etc.), (3) grouping by tens (40–, 50–, etc.):—

> 40, 43, 43, 46, 46, 46, 54, 56, 59, 62, 64, 64, 66, 66, 67, 67, 68, 68, 69, 69, 69, 71, 75, 75, 76, 76, 78, 80, 82, 82, 82, 82, 82, 83, 84, 86, 88, 90, 90, 91, 91, 92, 95, 102, 127.

8.9. Apply Sheppard's correction to the standard deviations calculated in Exercises 8.1 and 8.2 above.

8.10. (Continuing Exercise 7.9, p. 132.) Supposing the frequencies of values 0, 1, 2, 3, . . . of a variable to be given by the terms of the binomial series

$$q^n, \quad nq^{n-1}p, \quad \frac{n(n-1)}{1 \cdot 2}q^{n-2}p^2, \quad \ldots$$

where $p+q=1$, find the standard deviation.

8.11. (Cf. the remarks at the end of 8.20.) The sum of the deviations (without regard to sign) about the centre of the class-interval containing the mean (or median), in a grouped frequency-distribution, is found to be S. Find the correction to be applied to this sum, in order to reduce it to the mean (or median) as origin, on the assumption that the observations are evenly distributed over each class-interval. Take the number of observations below the interval containing the mean (or median) to be n_1, in that interval n_2 and above it n_3, and the distance of the mean (or median) from the arbitrary origin to be d.

8.12. (W. Scheibner, "Ueber Mittelwerthe," Berichte der kgl. sächsischen Gesellschaft d. Wissenschaften, 1873, p. 564, cited by Fechner, ref. (103): the second form of the relation is given by G. Duncker ("Die Methode der Variationsstatistik"; Leipzig, 1899) as an empirical one.) Show that if deviations are small compared with the mean, so that $(x/M)^3$ and higher powers of x/M may be neglected, we have approximately the relation

$$G = M\left(1 - \tfrac{1}{2}\frac{\sigma^2}{M^2}\right)$$

where G is the geometric mean, M the arithmetic mean and σ the standard deviation: and consequently to the same degree of approximation $M^2 - G^2 = \sigma^2$.

8.13. (Scheibner, loc. cit.) Similarly, show that if deviations are small compared with the mean, we have approximately

$$H = M\left(1 - \frac{\sigma^2}{M^2}\right)$$

H being the harmonic mean.

8.14. Find the coefficients of variation of the height distributions of Exercise 8.1 (using the uncorrected values of the s.d. as given).

8.15. Show that if a range of six times the standard deviation covers at least 18 class-intervals, Sheppard's correction will make a difference of less than 0·5 per cent. in the uncorrected value of the standard deviation.

MOMENTS AND MEASURES OF SKEWNESS
AND KURTOSIS.

Moments.

9.1. In considering the calculation of the mean and the root-mean-square deviation we have defined, in passing, the quantities $\frac{1}{N}S(f\xi)$ and $\frac{1}{N}S(f\xi^2)$ as the first and second moments about the value A, ξ being as before the value $X - A$, *i.e.* the excess of the variate value X over the value A. The first moment about the mean is zero, and the second moment about the mean is the variance (**8.5**).

In generalisation of these definitions we now define the *n*th **moment about A** as μ_n', where

$$\mu_n' = \frac{1}{N}S(f\xi^n) \qquad . \qquad . \qquad . \qquad . \qquad (9.1)$$

The **moments about the mean**, which are of particular importance, we write without dashes, so that

$$\mu_n = \frac{1}{N}S(fx^n) \qquad . \qquad . \qquad . \qquad . \qquad (9.2)$$

From these definitions we have :

$$\mu_0' = \mu_0 = \frac{1}{N}S(f) = 1 \qquad \text{since } \xi^0 \text{ and } x^0 = 1$$

$$\mu_1' = \frac{1}{N}S(f\xi) = M - A = d$$

$$\mu_1 = 0$$

$$\mu_2' = \frac{1}{N}S(f\xi^2) = \sigma^2 + d^2$$

$$\mu_2 = \sigma^2$$

These results we have already seen.

9.2. The word " moment " derives from Statics, and we may direct the attention of the student who is familiar with moments of forces to the fact that the sum $S(f\xi^n)$ is divided by N in the definition above. This amounts to a slight departure from the Statical practice, and some writers refer to what we have called " moments " as " moment-coefficients " in order to keep this fact in mind. In Statistics, however, no confusion is likely to arise from the use of the briefer form " moments."

The expression " moments " is also used by some writers to denote exclusively the moments about the mean, except in the case of the first

moment, which is zero about the mean, and which, therefore, is understood to be related to the origin under consideration at the moment. We shall not adopt this practice.

Moments about the Mean in terms of Moments about Any Point.

9.3.   We have, by definition,

$$\xi = X - A = (X - M) + (M - A)$$
$$= x + d$$

Hence,

$$f\xi^n = f(x + d)^n$$

and

$$S(f\xi^n) = S\{f(x + d)^n\}$$

Now, by the binomial theorem,

$$(x + d)^n = x^n + {}^nC_1 dx^{n-1} + {}^nC_2 d^2 v^{n-2} + \ldots + d^n$$

Hence,

$$S(f\xi^n) = S(fx^n) + {}^nC_1 dS(fx^{n-1}) + {}^nC_2 d^2 S(fx^{n-2}) + \ldots + d^n S(f)$$

Dividing by N we get :

$$\mu_n' = \mu_n + {}^nC_1 d\mu_{n-1} + {}^nC_2 d^2 \mu_{n-2} + \ldots + d^n \qquad . \qquad (9.3)$$

Similarly,

$$S(fx^n) = S\{f(\xi - d)^n\}$$

and

$$\mu_n = \mu_n' - {}^nC_1 d\mu'_{n-1} + {}^nC_2 d^2 \mu'_{n-2} - \ldots + (-1)^n d^n \qquad . \qquad (9.4)$$

These useful relations express the moments about the mean in terms of those about an arbitrary point A, and *vice versa*.

In particular we have :

If $n = 1$,

$$\mu_1' = \mu_1 + d = d \qquad\qquad \text{from (9.3)}$$
$$\mu_1 = \mu_1' - d = 0 \qquad\qquad \text{from (9.4)}$$

which are simply the relation $M - A = d$ in another form.

If $n = 2$,

$$\mu_2' = \mu_2 + 2d\mu_1 + d^2 \qquad\qquad \text{from (9.3)}$$
$$= \mu_2 + d^2 = \sigma^2 + d^2$$

$$\mu_2 = \mu_2' - 2d\mu_1' + d^2 \qquad\qquad \text{from (9.4)}$$
$$= \mu_2' - 2d^2 + d^2$$
$$= \mu_2' - d^2$$

These are the relation $\mu_2' = \sigma^2 + d^2$.

If $n = 3$,

$$\mu_3' = \mu_3 + 3d\mu_2 + 3d^2\mu_1 + d^3 \qquad\qquad \text{from (9.3)}$$
$$= \mu_3 + 3d\mu_2 + d^3 \qquad . \qquad . \qquad . \qquad . \qquad (9.5)$$

$$\mu_3 = \mu_3' - 3d\mu_2' + 3d^2\mu_1' - d^3 \qquad\qquad \text{from (9.4)}$$
$$= \mu_3' - 3d\mu_2' + 2d^3 \qquad . \qquad . \qquad . \qquad . \qquad (9.6)$$

If $n = 4$,

$$\mu_4' = \mu_4 + 4d\mu_3 + 6d^2\mu_2 + 4d^3\mu_1 + d^4 \qquad \text{from (9.3)}$$

$$= \mu_4 + 4d\mu_3 + 6d^2\mu_2 + d^4 \ . \qquad . \qquad . \qquad . \qquad (9.7)$$

$$\mu_4 = \mu_4' - 4d\mu_3' + 6d^2\mu_2' - 4d^3\mu_1' + d^4 \qquad \text{from (9.4)}$$

$$= \mu_4' - 4d\mu_3' + 6d^2\mu_2' - 3d^4 \ . \qquad . \qquad . \qquad . \qquad (9.8)$$

Calculation of Moments.

9.4. The calculation of moments of the third and higher orders is similar to that of the first and second. For grouped data we regard the observations as concentrated at the mid-points of the intervals ; we choose a convenient arbitrary origin A, find the moments about it and use the relations (9.3) and (9.4) above to find the moments about the mean ; we use a check on the arithmetic similar to that of **8.10** ; and we have under certain conditions certain Sheppard corrections for grouping.

In practice we rarely require to ascertain moments higher than the fourth. Indeed, moments of higher orders, though important in theory, are so extremely sensitive to sampling fluctuations that values calculated for moderate numbers of observations are quite unreliable and hardly ever repay the labour of computation.

9.5. There are various checks in use for the arithmetic of calculation. We shall use a generalisation of the simple identities of **7.12** and **8.10**. In fact, we have

$$(\xi + 1)^3 = \xi^3 + 3\xi^2 + 3\xi + 1$$

and hence,

$$S\{f(\xi + 1)^3\} = S(f\xi^3) + 3S(f\xi^2) + 3S(f\xi) + N$$

Similarly,

$$S\{f(\xi + 1)^4\} = S(f\xi^4) + 4S(f\xi^3) + 6S(f\xi^2) + 4S(f\xi) + N$$

and so on.

Thus, in calculating $S(f\xi^n)$ we also find $S\{f(\xi + 1)^n\}$, and this, together with the sums of lower orders, will give us a ready check on the work.

This check is sometimes known as the Charlier check, after C. V. L. Charlier, the Swedish Statistician.

Example 9.1. — Continuing our work on the height distribution of Table 6.7, page 94, let us find the third and fourth moments of the distribution about the mean.

In almost all practical work we require the first and second moments as a matter of course. It is therefore best to proceed systematically in the computation of the various moments by setting out the arithmetic in tabular form as on opposite page.

From this table we have :

$$S(f\xi) = 8{,}763 - 8{,}584 = 179$$
$$S(f\xi^2) = 56{,}809$$
$$S(f\xi^3) = 119{,}391 - 117{,}622 = 1{,}769$$
$$S(f\xi^4) = 1{,}182{,}061$$

CALCULATION OF FIRST FOUR MOMENTS *of the Distribution of Heights of Table 6.7, p. 94.*

Height, Inches.	f.	ξ.	$f\xi$.	$f(\xi+1)$.	$f\xi^2$.	$f(\xi+1)^2$.	$f\xi^3$.	$f(\xi+1)^3$.	$f\xi^4$.	$f(\xi+1)^4$.
57–	2	−10	− 20	− 18	200	162	− 2,000	− 1,458	20,000	13,122
58–	4	− 9	− 36	− 32	324	256	− 2,916	− 2,048	26,244	16,384
59–	14	− 8	− 112	− 98	896	686	− 7,168	− 4,802	57,344	33,614
60–	41	− 7	− 287	− 246	2,009	1,476	− 14,063	− 8,856	98,441	53,136
61–	83	− 6	− 498	− 415	2,988	2,075	− 17,928	−10,375	107,568	51,875
62–	169	− 5	− 845	− 676	4,225	2,704	− 21,125	−10,816	105,625	43,264
63–	394	− 4	−1,576	−1,182	6,304	3,546	− 25,216	−10,638	100,864	31,914
64–	669	− 3	−2,007	−1,338	6,021	2,676	− 18,063	− 5,352	54,189	10,704
65–	990	− 2	−1,980	− 990	3,960	990	− 7,920	− 990	15,840	·990
66–	1,223	− 1	−1,223	−4,995	1,223	—	1,223	−55,335	1,223	—
67–	1,329	0	−8,584	1,329	—	1,329	−117,622	1,329	—	1,329
68–	1,230	1	1,230	2,460	1,230	4,920	1,230	9,840	1,230	19,680
69–	1,063	2	2,126	3,189	4,252	9,567	8,504	28,701	17,008	86,103
70–	646	3	1,938	2,584	5,814	10,336	17,442	41,344	52,326	165,376
71–	392	4	1,568	1,960	6,272	9,800	25,068	49,000	100,352	245,000
72–	202	5	1,010	1,212	5,050	7,272	25,250	43,632	126,250	261,792
73–	79	6	474	553	2,844	3,871	17,064	27,097	102,384	189,679
74–	32	7	224	256	1,568	2,048	10,976	16,384	76,832	131,072
75–	16	8	128	144	1,024	1,296	8,192	11,664	65,536	104,976
76–	5	9	45	50	405	500	3,645	5,000	32,805	50,000
77–	2	10	20	22	200	242	2,000	2,662	20,000	29,282
Total	8,585	—	8,763	13,759	56,809	65,752	119,391	236,653	1,182,061	1,539,292

As a check on $S(f\xi^3)$ we have :

$$S(f\xi^3) + 3S(f\xi^2) + 3S(f\xi) + N$$
$$= 1,769 + 170,427 + 537 + 8,585$$
$$= 181,318$$
$$= S\{f(\xi+1)^3\}$$

As a check on $S(f\xi^4)$ we have :

$$S(f\xi^4) + 4S(f\xi^3) + 6S(f\xi^2) + 4S(f\xi) + N$$
$$= 1,182,061 + 7,076 + 340,854 + 716 + 8,585$$
$$= 1,539,292$$
$$= S\{f(\xi+1)^4\}$$

We have then :

$$d = \mu_1' = \frac{1}{N}S(f\xi) = \frac{179}{8,585} = \quad 0\cdot020,850,32$$

$$\mu_2' = \frac{56,809}{8,585} \qquad = \quad 6\cdot617,239,37$$

$$\mu_3' = \frac{1,769}{8,585} \qquad = \quad 0\cdot206,057,08$$

$$\mu_4' = \frac{1,182,061}{8,585} \qquad = 137\cdot689,108,91$$

$$\mu_2 = \mu_2' - d^2$$
$$= 6\cdot616,805$$

From equation (9.6):

$$\mu_3 = \mu_3' - 3d\mu_2' + 2d^3$$
$$= 0{\cdot}206{,}057{,}08 - 0{\cdot}413{,}914{,}67 + 0{\cdot}000{,}018{,}13$$
$$= -0{\cdot}207{,}839$$

From equation (9.8):

$$\mu_4 = \mu_4' - 4d\mu_3' + 6d^2\mu_2' - 3d^4$$
$$= 137{\cdot}689{,}108{,}91 - 0{\cdot}017{,}184{,}24 + 0{\cdot}017{,}260{,}51 - 0{\cdot}000{,}000{,}57$$
$$= 137{\cdot}689{,}185$$

which gives us μ_2, μ_3, μ_4 in units based on class-intervals, *i.e.* inches.

Example 9.2.—To find the moments about the mean of the distribution of Australian marriages of Table 6.8, page 96.

Until the last stage we work in class-intervals of 3 years. As in Example 8.3, page 140, we take a working mean at 28·5 years.

CALCULATION OF THE FIRST FOUR MOMENTS *of the Distribution of Marriages of Table 6.8, p. 96.*

Mid-value of Intervals, Years.	f.	ξ.	$f\xi$.	$f(\xi+1)$.	$f\xi^2$.	$f(\xi+1)^2$.	$f\xi^3$.	$f(\xi+1)^3$.	$f\xi^4$.	$f(\xi+1)^4$.
16·5	294	−4	− 1,176	− 882	4,704	2,646	− 18,816	− 7,938	75,264	23,814
19·5	10,995	−3	− 32,985	−21,990	98,955	43,980	−296,865	− 87,960	890,595	175,920
22·5	61,001	−2	−122,002	−61,001	244,004	61,001	−488,008	− 61,001	976,016	61,001
25·5	73,054	−1	− 73,054	−83,873	73,054	—	− 73,054	−156,899	73,054	—
28·5	56,501	0	−229,217	56,501	—	56,501	−876,743	56,501	.	56,501
31·5	33,478	1	33,478	66,956	33,478	133,912	33,478	267,824	33,478	535,648
34·5	20,569	2	41,138	61,707	82,276	185,121	164,552	555,363	329,104	1,666,089
37·5	14,281	3	42,843	57,124	128,529	228,496	385,587	913,984	1,156,761	3,655,936
40·5	9,320	4	37,280	46,600	149,120	233,000	596,480	1,165,000	2,385,920	5,825,000
43·5	6,236	5	31,180	37,416	155,900	224,496	779,500	1,346,976	3,897,500	8,081,856
46·5	4,770	6	28,620	33,390	171,720	233,730	1,030,320	1,636,110	6,181,920	11,452,770
49·5	3,620	7	25,340	28,960	177,380	231,680	1,241,660	1,853,440	8,691,620	14,827,520
52·5	2,190	8	17,520	19,710	140,160	177,390	1,121,280	1,596,510	8,970,240	14,368,590
55·5	1,655	9	14,895	16,550	134,055	165,500	1,206,495	1,655,000	10,858,455	16,550,000
58·5	1,100	10	11,000	12,100	110,000	133,100	1,100,000	1,464,100	11,000,000	16,105,100
61·5	810	11	8,910	9,720	98,010	116,640	1,078,110	1,399,680	11,859,210	16,796,160
64·5	649	12	7,788	8,437	93,456	109,681	1,121,472	1,425,853	13,457,664	18,536,089
67·5	487	13	6,331	6,818	82,303	95,452	1,069,939	1,336,328	13,909,207	18,708,592
70·5	326	14	4,564	4,890	63,896	73,350	894,544	1,100,250	12,523,616	16,503,750
73·5	211	15	3,165	3,376	47,475	54,016	712,125	864,256	10,681,875	13,828,096
76·5	119	16	1,904	2,023	30,464	34,391	487,424	584,647	7,798,784	9,938,999
79·5	73	17	1,241	1,314	21,097	23,652	358,649	425,736	6,097,033	7,663,248
82·5	27	18	486	513	8,748	9,747	157,464	185,193	2,834,352	3,518,667
85·5	14	19	266	280	5,054	5,600	96,026	112,000	1,824,494	2,240,000
88·5	5	20	100	105	2,000	2,205	40,000	46,305	800,000	972,405
Totals	301,785	—	318,049	474,490	2,155,838	2,635,287	13,675,105	19,991,056	137,306,162	202,091,751

From this table we have:

$$S(f\xi) = 318{,}049 - 229{,}217 = 88{,}832$$
$$S(f\xi^2) = 2{,}155{,}838$$
$$S(f\xi^3) = 13{,}675{,}105 - 876{,}743 = 12{,}798{,}362$$
$$S(f\xi^4) = 137{,}306{,}162$$

As a check on $S(f\xi)$ we have :

$$S(f\xi) + N = 88,832 + 301,785 = 390,617$$
$$= S\{f(\xi+1)\}$$

Similarly, for $S(f\xi^2)$:

$$S(f\xi^2) + 2S(f\xi) + N = 2,155,838 + 177,664 + 301,785$$
$$= 2,635,287$$
$$= S\{f(\xi+1)^2\}$$

As a check on $S(f\xi^3)$:

$$S(f\xi^3) + 3S(f\xi^2) + 3S(f\xi) + N$$
$$= 12,798,362 + 6,467,514 + 266,496 + 301,785$$
$$= 19,834,157$$
$$= S\{f(\xi+1)^3\}$$

As a check on $S(f\xi^4)$:

$$S(f\xi^4) + 4S(f\xi^3) + 6S(f\xi^2) + 4S(f\xi) + N$$
$$= 137,306,162 + 51,193,448 + 12,935,028 + 355,328 + 301,785$$
$$= 202,091,751$$
$$= S\{f(\xi+1)^4\}$$

Hence, about the working mean :

$$d = \mu_1' = \frac{88,832}{301,785} = 0 \cdot 294,355,253$$
$$\mu_2' = \frac{2,155,838}{301,785} = 7 \cdot 143,622,115$$
$$\mu_3' = \frac{12,798,362}{301,785} = 42 \cdot 408,873,867$$
$$\mu_4' = \frac{137,306,162}{301,785} = 454 \cdot 980,075,219$$

For moments about the mean :

$$\mu_2 = \mu_2' - d^2 = 7 \cdot 056,977$$
$$\mu_3 = \mu_3' - 3d\mu_2' + 2d^3 = 36 \cdot 151,595$$
$$\mu_4 = \mu_4' - 4d\mu_3' + 6d^2\mu_2' - 3d^4 = 408 \cdot 738,210$$

These are expressed in class-intervals, which are units of three years. If, as we rarely do, we wish to express the results in other units, say one year, we must multiply the first moment by 3, the second by 3^2, the third by 3^3, the fourth by 3^4, and so on ; *e.g.*

$$\mu_2 = 7 \cdot 056,977 \times 9 = 63 \cdot 512,79$$

In this and the preceding example we have retained more digits than are probably necessary, but the student will find it as well to retain several more than appear to be required, since subsequent work involving multiplication or addition may otherwise throw doubt on the final figures.

9.6. It will be evident that the labour involved in calculating the third and fourth moments is very considerable. Calculating machines

or tables of powers are a great help, and certain tables for the specific purpose of computing moments will be found in "*Tables for Statisticians and Biometricians, Part I*." The student should familiarise himself with the methods given in the two examples above, since, although we shall not use them to any great extent in this book, moments are important in more advanced theory.

Sheppard Corrections for Moments.

9.7. As in the case of the second moment, the effect due to grouping at mid-points of intervals may be corrected for by formulæ due to W. F. Sheppard, from whom they derive their name. The formulæ for the second, third and fourth moments are as follows :—

$$
\left.
\begin{aligned}
\mu_2' \text{ (corrected)} &= \mu_2 - \frac{h^2}{12} \\
\mu_3 \text{ (corrected)} &= \mu_3 \\
\mu_4 \text{ (corrected)} &= \mu_4 - \tfrac{1}{2}h^2\mu_2 + \frac{7}{240}h^4
\end{aligned}
\right\} \qquad . \qquad . \quad (9.9)
$$

where h is the width of the class-interval. If we are working in class-intervals as units, h is taken to be unity.

The use of these formulæ is restricted to the cases which we mentioned in **8.11**, *i.e.* those in which (*a*) the frequency-distribution is continuous, and (*b*) the distribution tapers off to zero in both directions.

. *Example 9.3.*—In Example 9.1 we found :

$$
\begin{aligned}
\mu_2 &= 6 \cdot 616,805 \\
\mu_3 &= -0 \cdot 207,839 \\
\mu_4 &= 137 \cdot 689,185
\end{aligned}
$$

Applying the above corrections, h being 1 :

$$
\begin{aligned}
\mu_2 \text{ (corr.)} &= 6 \cdot 616,805 - 0 \cdot 083,333 \\
&= 6 \cdot 533,472 \\
\mu_3 \text{ (corr.)} &= -0 \cdot 207,839 \\
\mu_4 \text{ (corr.)} &= 137 \cdot 689,185 - 3 \cdot 308,402 + 0 \cdot 029,167 \\
&= 134 \cdot 409,950
\end{aligned}
$$

Example 9.4.—In Example 9.2 we have, in units of 3 years :

$$
\begin{aligned}
\mu_2 &= 7 \cdot 056,977 \\
\mu_3 &= 36 \cdot 151,595 \\
\mu_4 &= 408 \cdot 738,21
\end{aligned}
$$

Thus :

$$
\begin{aligned}
\mu_2 \text{ (corr.)} &= 7 \cdot 056,977 - 0 \cdot 083,333 \\
&= 6 \cdot 973,644 \\
\mu_3 \text{ (corr.)} &= 36 \cdot 151,595 \\
\mu_4 \text{ (corr.)} &= 408 \cdot 738,210 - 3 \cdot 528,489 + 0 \cdot 029,167 \\
&= 405 \cdot 238,888 \quad .
\end{aligned}
$$

In units of one year the corrected moments are given by multiplying by **9, 27** and **81** as before.

β- and γ-Coefficients.

9.8. Certain quantities calculated from the moments about the mean are of particular importance in statistical work. We define

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \qquad . \qquad . \qquad . \qquad . \qquad (9.10)$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \qquad . \qquad . \qquad . \qquad . \qquad (9.11)$$

and two further quantities :

$$\gamma_1 = +\sqrt{\beta_1} \qquad . \qquad . \qquad . \qquad . \qquad (9.12)$$

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4 - 3\mu_2^2}{\mu_2^2} \qquad . \qquad . \qquad . \qquad (9.13)$$

The reason for the introduction of these arbitrary-looking quantities will appear in the sequel.[1]

It is to be noted that these four coefficients are all pure numbers and, as such, are independent of the scale of measurement of the variable ; for since μ_n has the dimensions of (variable)^n, μ_3^2 has the dimensions (variable)^6 and so has μ_2^3, and hence their quotient has dimension zero, i.e. is a pure number ; and similarly for the quotient of μ_4 and μ_2^2.

Example 9.5.—Let us calculate β_1 and β_2 for the distribution of Example 9.1.

We have, using the corrected values of Example 9.3 :

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$= \frac{(-0\cdot207839)^2}{(6\cdot533472)^3}$$

$$= \frac{0\cdot043197}{278\cdot889} = 0\cdot000155$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$= \frac{134\cdot40995}{42\cdot68662}$$

$$= 3\cdot149$$

Example 9.6.—Similarly, in the data of Example 9.2, using corrected values :

$$\beta_1 = \frac{(36\cdot151595)^2}{(6\cdot973644)^3}$$

$$= 3\cdot854$$

$$\beta_2 = \frac{405\cdot238888}{(6\cdot973644)^2}$$

$$= 8\cdot333$$

---

[1] In general, Karl Pearson defines

$$\beta_{2n+1} = \frac{\mu_3 \mu_{2n+3}}{\mu_2^{n+3}}$$

$$\beta_{2n} = \frac{\mu_{2n+2}}{\mu_2^{n+1}}$$

It should be noted in this last example that, since the coefficients are pure numbers, it does not matter whether we work in units of three years or of one year.

Measures of Skewness.

9.9. The departure of a frequency-distribution from symmetry has a certain interest, and several measures have been devised to permit of the measurement of this skewness. Such measures should (*a*) be pure numbers, so as to be independent of the units in which the variable is measured, and (*b*) be zero when the distribution is symmetrical.

9.10. Three such measures deserve mention. In the first place, we can define

$$\text{Skewness} = \frac{(Q_3 - Mi) - (Mi - Q_1)}{2Q} = \frac{Q_1 + Q_3 - 2Mi}{2Q} \quad . \quad (9.14)$$

·This can be put in the form :

$$\text{Skewness} = \frac{(Q_3 - Mi) - (Mi - Q_1)}{(Q_3 - Mi) + (Mi - Q_1)} \quad . \quad . \quad (9.15)$$

i.e. the skewness is taken to be the difference of the quartile deviations from the median divided by their sum. It is clearly a pure number, for both numerator and denominator have the same dimensions, and it is zero when the distribution is symmetrical. It varies from -1 to $+1$.[1]

This is a rather rough-and-ready measure which might, however, be useful if we were using the semi-interquartile range as a measure of dispersion and were unable or unwilling to calculate the standard deviation.

9.11. The most common measure of skewness is Pearson's, defined by

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} = \frac{M - Mo}{\sigma} \quad . \quad . \quad (9.16)$$

This evidently is a pure number and is zero for symmetrical distributions.

9.12. The calculation of this coefficient of skewness is subject to the inconvenience of determining the position of the mode. We may circumvent this difficulty in several ways. In the first place, for distributions which are obviously not too skew we may use the empirical relation of **7.27.** We then have :

$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}} \quad . \quad . \quad (9.17)$$

Secondly, for a large class of curves to which the moderately skew humped curve is a close approximation, the skewness of equation (9.16) is given exactly by

$$\text{Skewness} = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)} \quad . \quad . \quad (9.18)$$

---

[1] In the 10th and previous editions of this book the measure Skewness $= \dfrac{Q_1 + Q_3 - 2Mi}{Q}$ was suggested, *i.e.* twice the measure (9.14). The above form has the advantage that its limits are -1 and $+1$.

We may, therefore, take this to be an approximation to the value given by equation (9.16).

It should be noted that the measures (9.14) and (9.16) are positive if the longer tail of the distribution lies toward the higher values of the variate (the right) and negative in the contrary case. This accords with the anticipatory remarks of **6.20**. The measure (9.18) is to be regarded as without sign.

Limits of the Measures of Skewness.

9.13. We have already remarked that the measure given by equation (9.14) lies between -1 and $+1$. There is no limit in theory to the measure (9.16) or its approximation (9.18), and this is a slight drawback. But in practice the value given by equation (9.16) is rarely very high, and for moderately skew single-humped curves is usually less than unity.

It has been shown that the quantity $\dfrac{\text{Mean} - \text{Median}}{\text{Standard deviation}}$ lies between the limits -1 and $+1$, and the measure (9.17) therefore lies between -3 and $+3$ (see ref. (161)). In practice it rarely approaches these limits.

Example 9.7.—Let us once again consider the height distribution of Table 6.7, which has been already discussed in this chapter (Examples 9.1, 9.3 and 9.5).

We have :

Mean (Example 7.1, p. 116)	$= 67{\cdot}46$ inches
S.d. (corrected, Example 8.4, p. 141)	$= 2{\cdot}56$ inches
Median (Example 7.3, p. 121)	$= 67{\cdot}47$ inches
Q_1 (Example 8.9, p. 148)	$= 65{\cdot}71$ inches
Q_3 (*ibid.*)	$= 69{\cdot}21$ inches
Q (*ibid.*)	$= 1{\cdot}75$ inches
β_1 (corrected, Example 9.5, p. 161)	$= 0{\cdot}000155$
β_2 (*ibid.*)	$= 3{\cdot}149$

The measure of skewness (9.14) is, then,

$$\text{Sk} = \frac{Q_1 + Q_3 - 2Mi}{2Q}$$
$$= \frac{65{\cdot}71 + 69{\cdot}21 - (2 \times 67{\cdot}47)}{2 \times 1{\cdot}75}$$
$$= -0{\cdot}006$$

We can clearly place no reliance on this figure. The median and quartiles were obtained by methods of approximation which we cannot expect to give accuracy to the second decimal place. We can only conclude, therefore, that so far as the measure (9.14) is concerned, there is no significant skewness.

The measure (9.18) gives :

$$\text{Sk} = \frac{0{\cdot}0124 \times 6{\cdot}149}{2(15{\cdot}745 - 0{\cdot}001 - 9)}$$
$$= \frac{0{\cdot}0124 \times 6{\cdot}149}{2 \times 6{\cdot}744}$$
$$= 0{\cdot}006$$

Here again the skewness is extremely small, and is, in fact, almost equal to the value given by (9.14).

If we take the measure (9.17) we get:

$$Sk = \frac{3(M - Mi)}{\sigma}$$

$$= \frac{-0.03}{2.56}$$

$$= -0.012$$

This value is suspect because we have determined the mean and the median only to the second decimal place, but clearly the value is small.

We conclude that there is only very slight skewness. At this stage we cannot say whether such small skewness is significant, but it is at least probably attributable to sampling fluctuations

Example 9.8.—For the marriage data of Examples 9.2, 9.4 and 9.6 it will be found that, using the working mean as origin :

$$Mean = \quad 0.2944$$
$$Median = -0.4018$$
$$Q_1 = -1.4568$$
$$Q_3 = \quad 1.2316$$

and

$$\sigma \text{ (corrected) (Ex. 8.5)} = 2.6408$$
$$\beta_1 = 3.854$$
$$\beta_2 = 8.333$$

The measure (9.14) is :

$$Sk = \frac{(Q_3 - Mi) - (Mi - Q_1)}{(Q_3 - Mi) + (Mi - Q_1)}$$

$$= \frac{1.6334 - 1.0550}{1.6334 + 1.0550}$$

$$= \frac{0.5784}{2.6884}$$

$$= 0.22$$

The measure (9.18) is :

$$Sk = \frac{\sqrt{3.854}(11.333)}{2(41.665 - 23.124 - 9)}$$

$$= \frac{1.963 \times 11.333}{2 \times 9.541}$$

$$= 1.17$$

The two are very different, as we might expect, but both indicate strong positive skewness. As a matter of interest we may compare the value (9.17), which gives

$$Sk = \frac{3 \times 0.6962}{2.6408}$$

$$= 0.79$$

Kurtosis.

9.14. The coefficient β_2 or its derivative γ_2 is used to measure a property of the single-humped distribution known as **kurtosis** ($\kappa\nu\rho\tau\acute{o}s$, humped).

We take as the standard value of β_2 the number 3, for reasons which will appear when we study the so-called "normal" curve **(10.24)**. This curve is approximately of the shape given in fig. 6.5, page 93. Curves with values of β_2 less than 3 will, compared with this, be flat-topped, and are called **platykurtic** ($\pi\lambda\check{\alpha}\tau\acute{\nu}s$, broad, $+\kappa\nu\rho\tau\acute{o}s$). Curves with values greater than 3 will be peaked more sharply, and are called **leptokurtic**[1] ($\lambda\epsilon\pi\tau\acute{o}s$, narrow, $+\kappa\nu\rho\tau\acute{o}s$). "Student" gives an amusing mnemonic for these names : Platykurtic curves, like the platypus, are squat with short tails. Leptokurtic curves are high with long tails like the kangaroo—noted for " lepping " !

Example 9.9.—In the height distribution of Examples 9.1, 9.3, 9.5 and 9.7 :

$$\beta_2 = 3\cdot149$$
$$\gamma_2 = \beta_2 - 3 = 0\cdot149$$

Hence the curve is slightly leptokurtic.

On the other hand, in the marriage distribution of Examples 9.2, 9.4, 9.6 and 9.8 :

$$\beta_2 = 8\cdot333$$
$$\gamma_2 = 5\cdot333$$

and the curve is very leptokurtic.

Seminvariants.

9.15. We may conclude this chapter by referring briefly to a set of quantities similar to moments which have some theoretical and practical importance. These are Thiele's **seminvariants**.

The seminvariants are defined by a rather complicated mathematical expression which we shall not here reproduce. For present purposes it is sufficient to note that the first four seminvariants may be expressed as simple functions of the first four moments. In fact we have :

$$\left.\begin{aligned}
\lambda_1 &= \mu_1' \\
\lambda_2 &= \mu_2' - \mu_1'^2 \\
\lambda_3 &= \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 \\
\lambda_4 &= \mu_4' - 4\mu_1'\mu_3' - 3\mu_2'^2 + 12\mu_1'^2\mu_2' - 6\mu_1'^4
\end{aligned}\right\} \qquad . \qquad (9.19)$$

In particular, about the mean,

$$\left.\begin{aligned}
\lambda_1 &= 0 \\
\lambda_2 &= \mu_2 \\
\lambda_3 &= \mu_3 \\
\lambda_4 &= \mu_4 - 3\mu_2^2
\end{aligned}\right\} \qquad . \qquad . \qquad . \qquad (9.20)$$

---

[1] These terms are due to Karl Pearson and appear to have been given for the first time in *Biometrika*, vol. 4, 1905–6, page 169 *et seq.* By a slip *leptokurtosis* is there inadvertently applied to distributions for which $\beta_2 < 3$.

9.16. These relations are used in the calculation of the seminvariants, the moments being first ascertained in the manner of the earlier sections of this chapter. For instance, the first four seminvariants of the height distribution which has served us as an example are, about the mean,

$$\lambda_1 = 0$$
$$\lambda_2 = 6 \cdot 616805$$
$$\lambda_3 = -0 \cdot 207839$$
$$\lambda_4 = 137 \cdot 689185 - 3 \times (6 \cdot 616805)^2 = 6 \cdot 34286$$

if we take uncorrected values of the moments.

9.17. The seminvariants owe their name to two very remarkable properties. In the first place, all seminvariants except the first are independent of the origin of calculation. The moments vary according to the point about which they are calculated, which makes it necessary to specify the origin A in speaking of them. The seminvariants, on the other hand, do not, so that it is unnecessary to specify any value A in giving their values; the sole exception to this rule is the first seminvariant, which is the same as the first moment.

Secondly, if the scale of measurement of the variate is altered by multiplying all values by a constant a, the nth seminvariant is multiplied by a^n. Thus, in the height distribution, if we change our scale to centimetres instead of inches, and so multiply all values of the variate by 2·54, the seminvariants in the previous section are to be multiplied by 2·54, 2·54^2, 2·54^3, 2·54^4, respectively.

SUMMARY.

1. The nth moment about the point A is defined as

$$\mu_n' = \frac{1}{N}S(f\xi^n)$$

where $\xi = X - A$, and X is the value of the variate.

2. The nth moment about the mean is written μ_n.

3. $$\mu_n = \mu_n' - {}^nC_1 d\mu_{n-1}' + {}^nC_2 d^2\mu_{n-2}' - \ldots + (-1)^{n+1}d^n$$

where

$$d = M - A$$

and in particular

$$\mu_3 = \mu_3' - 3d\mu_2' + 2d^3$$
$$\mu_4 = \mu_4' - 4d\mu_3' + 6d^2\mu_2' - 3d^4$$

4. Sheppard's corrections for the moments are :

$$\mu_2 \text{ (corrected)} = \mu_2 - \frac{h^2}{12}$$
$$\mu_3 \text{ (corrected)} = \mu_3$$
$$\mu_4 \text{ (corrected)} = \mu_4 - \tfrac{1}{2}h^2\mu_2 + \frac{7}{240}h^4$$

5.
$$\beta_1 = \frac{\mu_3{}^2}{\mu_2{}^3} \qquad \beta_2 = \frac{\mu_4}{\mu_2{}^2}$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2{}^{\frac{3}{2}}} \qquad \gamma_2 = \beta_2 - 3 = \frac{\mu_4 - 3\mu_2{}^2}{\mu_2{}^2}$$

6. Pearson's measure of skewness is given by

$$Sk = \frac{Mean - Mode}{Standard\ deviation}$$

which, for a large class of curves, is equal to

$$\frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

7. If the standard deviation is not known, a rough measure of skewness is obtained by taking

$$Sk = \frac{Q_1 + Q_3 - 2Mi}{2Q}$$

8. Distributions for which $\beta_2 > 3$ are said to be leptokurtic ; those for which $\beta_2 < 3$ are platykurtic.

9. The first four seminvariants, in terms of the moments about the mean, are :

$$\lambda_1 = 0$$
$$\lambda_2 = \mu_2$$
$$\lambda_3 = \mu_3$$
$$\lambda_4 = \mu_4 - 3\mu_2{}^2$$

10. The seminvariants are independent of the origin of calculation, except the first, which is equal to the mean.

EXERCISES.

9.1. Find the first four moments about the mean of the distribution of males in the United Kingdom according to weight given in Exercise 6.6, page 111. (Correct your values for grouping.)

Hence find β_1 and β_2 and measure the kurtosis of the distribution.

9.2. For the same distribution find the three measures of skewness, approximating to the mode by the empirical relation of 7.27.

9.3. Find the first four moments about the mean, the values of β_1, β_2, and the three measures of skewness for the following distribution (see table, p. 168). (Apply Sheppard's corrections.)

9.4. In the data of Example 9.1, group the individuals by intervals of three inches (57–, 60–, etc.) and calculate the first four moments about the mean. Compare your results with those of Example 9.1, (a) before Sheppard's corrections are applied, and (b) after Sheppard's corrections are applied.

9.5. Find the third and fourth moments about the mean of the binomial series :

$$q^n, \qquad nq^{n-1}p, \qquad \frac{n(n-1)}{1 \cdot 2}q^{n-2}p^2, \quad \ldots \text{ where } p + q = 1$$

(continuing the work of Exercise 8.10, p. 153).

4912 Cows Classified according to their Yield of Milk. (Data from J. F. Tocher, "An Investigation of the Milk Yield of Dairy Cows," *Biometrika,* vol. 20B, 1928, pp. 105–244.)

Yield of Milk (gallons per week). (Central Value of Interval.)	Number of Cows.	Yield of Milk (gallons per week). (Central Value of Interval.)	Number of Cows.
8	1	23	214
9	5	24	153
10	13	25	112
11	33	26	58
12	71	27	35
13	151	28	13
14	236	29	15
15	339	30	4
16	499	31	5
17	552	32	2
18	585	33	1
19	586	34	1
20	496		
21	448	Total	4912
22	284		

9.6. The first four moments of a distribution about the value 4 are $-1 \cdot 5$, 17, -30 and 108; find the moments about the mean and the origin.

9.7. Show that for a symmetrical distribution all moments about the mean of odd order are zero.

9.8. Show that for any distribution $\beta_2 > 1$.

9.9. Calculate the second, third and fourth seminvariants of the distribution of Australian marriages of Example 9.2, (*a*) from the moments about the mean, using equation (9.20), and (*b*) from the moments about the value 28·5, using equation (9.19); and hence verify that the values of the seminvariants are independent of the origin of calculation. (Use uncorrected values of the moments.)

9.10. Show that

$$d = \lambda_1$$
$$\sigma^2 = \sqrt{\lambda_2}$$
$$\gamma_1 = \frac{\lambda_3}{\sigma^3}$$
$$\gamma_2 = \frac{\lambda_4}{\lambda_2^2}$$

THREE IMPORTANT THEORETICAL DISTRIBUTIONS — THE BINOMIAL, THE NORMAL AND THE POISSON.

Theoretical Distributions.

10.1. In the examples of frequency-distributions which we have given in Chapter 6 and subsequent chapters we have been careful to take data from observation and experiment. It is possible, however, starting with certain general hypotheses, to deduce mathematically what the frequency-distributions of certain universes should be. Such distributions we shall call **theoretical**.

10.2. There are three theoretical distributions which, from their historical interest as well as their intrinsic importance, occupy a position in the forefront of statistical theory. They are, in the order of their discovery, the **Binomial** (due to James Bernoulli, *circa* 1700), the **Normal** (due to Demoivre, but more often associated with the names of Laplace and Gauss, who discussed it at the close of the eighteenth and the beginning of the nineteenth centuries), and the **Poisson** (due to S. D. Poisson, who published it in 1837).

These three are, so to speak, the classical distributions. Certain others were discovered during the nineteenth century, but it was not until the end of the century that there began the second period of statistical discovery which has since given us a wealth of theoretical distributions. Even this latest crop depends to some extent on the properties of the first three, and particularly of the Normal Distribution. The three therefore form, historically and logically, the starting-point of the theory of particular distributions, and in this chapter we propose to give an account of their main properties.

The Binomial Distribution.

10.3. If we may regard an ideal coin as a uniform, homogeneous circular disc, there is nothing which can make it tend to fall more often on the one side than on the other; we may expect, therefore, that in any long series of throws the coin will fall with either face uppermost an approximately equal number of times, or with, say, heads uppermost approximately half the times. Similarly, if we may regard the ideal die as a perfect homogeneous cube, it will tend, in any long series of throws, to fall with each of its six faces uppermost an approximately equal number of times, or with any given face uppermost one-sixth of the whole number of times. These results are sometimes expressed by saying that the chance of throwing heads (or tails) with a coin is 1/2, and the chance of throwing six (or any other face) with a die is 1/6. To avoid speaking of such particular instances as coins or dice, we shall in future, using terms which

have become conventional, refer to an **event** the chance of **success** of which is p and the chance of **failure** q. Obviously $p + q = 1$.

10.4. We will now assume that the events in a number of trials are all independent, *i.e.* that the chances p and q are the same for each event and remain constant throughout the trials. The case corresponds to the tossing of perfect coins or the throwing of perfect dice.

Suppose now we take a number of sets of n trials and count the number of successes in each set ; for example, we might toss a coin ten times for each set, and observe the number of heads in each set of ten. In general, there will be some sets with no successes, some with one success, some with two successes, and so on. Hence, if we classify the sets according to the number of successes which they contain we shall get a frequency-distribution. Table 6.15, page 107, gives such a distribution for some dice-throwing experiments. We shall now see how, on the assumption of independence of successive events to which we have just referred, the nature of this distribution may be theoretically determined.

10.5. For the case of single events we expect in N trials to get Np successes and Nq failures.

Suppose now we take N pairs of events, *i.e.* two to the set. There will be Nq cases in which the first event is a failure, and, in virtue of the independence of the events, among these Nq there will be $Nq \times q$ failures, and $Nq \times p$ successes, of the second event on the average. Similarly, of the Np cases in which the first event was a success, the second event will, on the average, be a success in $Np \times p$ and a failure in $Np \times q$ cases. Hence there will be Nq^2 cases in which both events are failures, $2Npq$ cases with one success and one failure, and Np^2 cases in which both are successes.

If we now take N sets of three events, we see that, of the Nq^2 cases in which the first two events were failures, $Nq^2 \times q$ will give a third failure and $Nq^2 \times p$ one success ; of the $2Npq$ cases, $2Npq^2$ will give two failures and a success and $2Np^2q$ one failure and two successes ; and of the Np^2 cases, Np^2q will give one failure and two successes and Np^3 will give three successes. Hence the number of sets with 3 failures, 2 failures and 1 success, 1 failure and 2 successes, and 3 successes are, respectively,

$$Nq^3, \qquad 3Nq^2p, \qquad 3Nqp^2, \qquad Np^3$$

10.6. From these results it is evident that the frequencies of 0, 1, 2, . . . successes are given

for *one* event by the binomial expansion of $N(q + p)$
for *two* events   ,,   ,,   ,, $N(q + p)^2$
for *three* events   ,,   ,,   ,, $N(q + p)^3$

In general, for n events the frequencies of successes in N sets are given by the successive terms in the binomial expansion of $N(q + p)^n$, *i.e.*

$$N\left\{ q^n + nq^{n-1}p + \frac{n(n-1)}{1 \cdot 2}q^{n-2}p^2 + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3}q^{n-3}p^3 + \ . \ . \ . \right\}$$

This is the so-called **Binomial Distribution.**

Example 10.1.—If we take 100 sets of 10 tosses of a perfect coin, in how many cases should we expect to get 7 heads and 3 tails ?

Here $p = \tfrac{1}{2}, \qquad q = \tfrac{1}{2}$

Hence, the numbers of successes 0, 1, . . . 10 are the terms in $100(\frac{1}{2} + \frac{1}{2})^{10}$,

i.e. $100\left\{\left(\dfrac{1}{2}\right)^{10} + 10 \cdot \left(\dfrac{1}{2}\right)^9\left(\dfrac{1}{2}\right) + \dfrac{10 \cdot 9}{1 \cdot 2}\left(\dfrac{1}{2}\right)^8\left(\dfrac{1}{2}\right)^2 + \ldots \right\}$

The term giving 7 successes and 3 failures is :

$$100 \times {}^{10}C_7(\tfrac{1}{2})^7(\tfrac{1}{2})^3$$

$$= 100 \cdot \frac{10 \cdot 9 \cdot 8}{1 \cdot 2 \cdot 3} \cdot \frac{1}{2^{10}}$$

$$= \frac{3000}{256}$$

$$= 12 \text{ approximately}$$

Example 10.2.—In the previous example, in how many cases should we expect to get 7 heads at least ?   As before, the numbers of successes are the terms in

$$\frac{100}{2^{10}}\left\{1 + 10 + \frac{10 \cdot 9}{1 \cdot 2} + \ldots\right\}$$

We require the *sum* of terms with 7, 8, 9, 10 successes.   Our expected number is, then,

$$\frac{100}{2^{10}}\{{}^{10}C_7 + {}^{10}C_8 + {}^{10}C_9 + {}^{10}C_{10}\}$$

$$= \frac{100}{2^{10}}\left\{\frac{10 \cdot 9 \cdot 8}{1 \cdot 2 \cdot 3} + \frac{10 \cdot 9}{1 \cdot 2} + \frac{10}{1} + 1\right\}$$

$$= \frac{100}{2^{10}}\{176\}$$

$$= \frac{1100}{64}$$

$$= 17 \text{ approximately}$$

General Form of the Binomial Distribution.

10.7.   The form of the binomial distribution depends (1) on the values of p and q, (2) on the value of the exponent n.

If p and q are equal the distribution is evidently symmetrical, for p and q may be interchanged without altering the value of any term, and consequently terms equidistant from the two ends of the series are equal.
  • If, on the other hand, p and q are unequal, the distribution is skew. The following table shows the calculated distributions for $n = 20$ and values of p, proceeding by 0·1, from 0·1 to 0·5.   When $p = 0·1$, cases of two successes are the most frequent, but cases of one success almost equally frequent : even nine successes may, however, occur about once in 10,000 trials.   As p is increased, the position of the maximum frequency gradually advances, and the two tails of the distribution become more nearly equal, until $p = 0·5$, when the distribution is symmetrical.   Of course, if the table were continued, the distribution for $p = 0·6$ would be similar to that for $q = 0·6$, but reversed end for end, and so on.

TABLE 10.1.—*Terms of the Binomial Series 10,000 $(q+p)^{20}$ for Values of p from 0·1 to 0·5. (Figures given to the nearest unit.)*

Number of Successes.	$p=0·1$ $q=0·9$	$p=0·2$ $q=0·8$	$p=0·3$ $q=0·7$	$p=0·4$ $q=0·6$	$p=0·5$ $q=0·5$
0	1216	115	8	—	—
1	2702	576	68	5	—
2	2852	1369	278	31	2
3	1901	2054	716	123	11
4	898	2182	1304	350	46
5	319	1746	1789	746	148
6	89	1091	1916	1244	370
7	20	545	1643	1659	739
8	4	222	1144	1797	1201
9	1	74	654	1597	1602
10	—	20	308	1171	1762
11	—	5	120	710	1602
12	—	1	39	355	1201
13	—	—	10	146	739
14	—	—	2	49	370
15	—	—	—	13	148
16	—	—	—	3	46
17	—	—	—	—	11
18	—	—	—	—	2
19	—	—	—	—	—
20	—	—	—	—	—

10.8. If $p=q$, the effect of increasing n is to raise the mean and increase the dispersion. If p is not equal to q, however, not only does an increase in n raise the mean and increase the dispersion, but it also lessens the asymmetry; the greater n, for the same values of p and q, the less the asymmetry. Thus, if we compare the first distribution of the above table with that given by $n=100$, we have the following :—

TABLE 10.2.—*Terms of the Binomial Series 10,000 $(0·9 + 0·1)^{100}$. (Figures given to the nearest unit.)*

Number of Successes.	Frequency.	Number of Successes.	Frequency.	Number of Successes.	Frequency.
0	—	8	1148	16	193
1	3	9	1304	17	106
2	16	10	1319	18	54
3	59	11	1199	19	26
4	159	12	988	20	12
5	339	13	743	21	5
6	596	14	513	22	2
7	889	15	327	23	1

The maximum frequencies now occur for 9 and 10 successes, and the two " tails " are much more nearly equal. If, on the other hand, n is reduced to 2, the distribution is :

Number of Successes.	Frequency.
0	8100
1	1800
2	100

and the maximum frequency is at one end of the range.

The tendency towards symmetry may be seen from fig. 10.1, in which
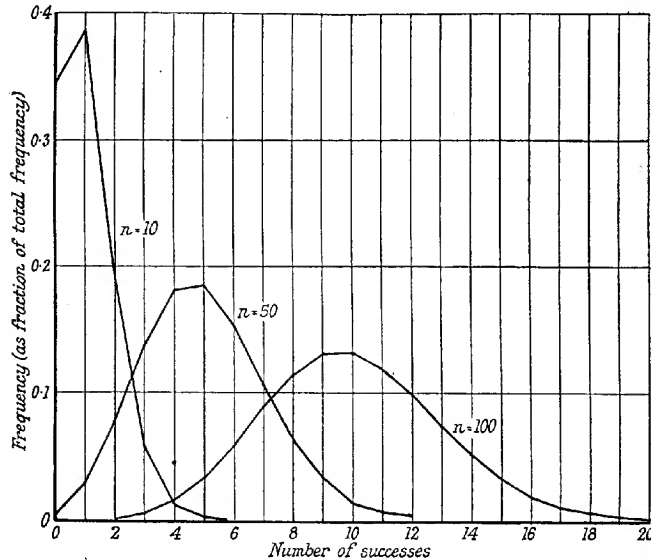
Fig. 10.1.—Frequency-polygons of the Binomial $(0·9 + 0·1)^n$ for Various Values of n.

the binomial $(0·9 + 0·1)^n$ has been drawn for various values of n. See also **10.12** below.

Constants of the Binomial Distribution.

10.9. We proceed to find the lower moments of the distribution $N(q+p)^n$.

Taking an arbitrary origin at 0 successes, we have the successive deviations ξ as $0, 1, 2, \ldots n$, and hence,

$$\mu_1' = (q^n \times 0) + ({}^nC_1 q^{n-1} p \times 1) + ({}^nC_2 q^{n-2} p^2 \times 2) + \ldots + (p^n \times n)$$
$$= p\{nq^{n-1} + n(n-1)q^{n-2}p + \ldots + np^{n-1}\}$$
$$= np\{q^{n-1} + (n-1)q^{n-2}p + \ldots + p^{n-1}\}$$
$$= np(q+p)^{n-1}$$

Now, $q + p = 1$

Hence, $\mu_1' = np$

That is, the mean M is np.

We have, further,

$$\mu_2' = (q^n \times 0) + ({}^nC_1 q^{n-1} p \times 1) + ({}^nC_2 q^{n-2} p^2 \times 2^2) + \ldots + (p^n \times n^2)$$

$$= np\{q^{n-1} + 2(n-1)q^{n-2}p + \frac{3(n-1)(n-2)}{2} q^{n-3}p^2 + \ldots + np^{n-1}\}$$

The expression in brackets is the first moment of the binomial $(q+p)^{n-1}$ about origin -1, and hence is equal to $(n-1)p + 1$.

Hence,

$$\mu_2' = np\{(n-1)p + 1\}$$

It may also be shown in a similar way (but we omit the proof) that

$$\mu_3' = np\{(n-1)(n-2)p^2 + 3(n-1)p + 1\}$$
$$\mu_4' = np\{(n-1)(n-2)(n-3)p^3 + 6(n-1)(n-2)p^2 + 7(n-1)p + 1\}$$

10.10.   From these results we may find the moments about the mean. We have :

$$\mu_2 = \mu_2' - d^2$$
$$= np\{(n-1)p + 1\} - n^2 p^2$$
$$= np(1-p)$$
$$= npq$$

Hence we have the important result that

$$\sigma = \sqrt{npq} \ . \qquad . \qquad . \qquad . \qquad . \qquad (10.1)$$

10.11.   Similarly, it will be found that

$$\mu_3 = npq(q-p) \qquad . \qquad . \qquad . \qquad . \qquad (10.2)$$

$$\mu_4 = 3p^2q^2n^2 + pqn(1-6pq) \qquad . \qquad . \qquad (10.3)$$

Hence,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(q-p)^2}{npq} \qquad . \qquad . \qquad . \qquad . \qquad (10.4)$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{1-6pq}{pqn} \qquad . \qquad . \qquad . \qquad (10.5)$$

10.12.   Thus the binomial distribution has mean np and standard deviation $\sqrt{npq}$.   It is instructive to note that β_1 and $(\beta_2 - 3)$ are both of order $\frac{1}{n}$.   Hence, as n becomes larger, the distribution tends to symmetry and zero kurtosis.

The values of β_1 and β_2 for some values of p and q and ranges of n are shown in Tables 10.3, 10.4 and 10.5.

From an inspection of these tables it will be seen that even for an extremely small value of p the binomial tends to zero β_1 and zero kurtosis for values of n well within practical limits.   For the symmetrical binomial $p = q = 0.5$, β_1 is of course zero, and β_2 rapidly approaches 3.

TABLE 10.3.—*Values of β_1 and β_2 for the Binomial with $p = 0.02$, $q = 0.98$.*
(From M. Greenwood, *Biometrika*, vol. 9, 1913, p. 69.)

n.	β_1.	β_2.
100	0·4702	3·4502
200	0·2351	3·2251
300	0·1567	3·1501
400	0·1176	3·1126
500	0·0940	3·0900
600	0·0784	3·0750
700	0·0672	3·0643
800	0·0588	3·0563
900	0·0522	3·0500
1000	0·0470	3·0450

TABLE 10.4.—*Values of β_1 and β_2 for the Binomial with $p = 0.1$, $q = 0.9$.*

n.	β_1.	β_2.
100	0·0711	3·0511
200	0·0356	3·0256
1000	0·0071	3·0051

TABLE 10.5.—*Values of β_2 for the Binomial with $p = 0.5$, $q = 0.5$.*

n.	β_2.
4	2·5
6	2·6667
8	2·75
10	2·8
50	2·96
100	2·98
1000	2·998

Mechanical Representation of the Binomial Distribution.

10.13. There is an interesting mechanical method of constructing a representation of the binomial series. The apparatus, which is illustrated in fig. 10.2, consists of a funnel opening into a space—say a $\frac{1}{4}$ inch in depth —between a sheet of glass and a back-board. This space is broken up by successive rows of wedges like 1, 2 3, 4 5 6, etc., which will divide up into streams any granular material such as shot or mustard seed which is poured through the funnel when the apparatus is held at a slope. At the foot, these wedges are replaced by vertical strips, in the spaces between which the material can collect. Consider the stream of material that comes from the funnel and meets the wedge 1. This wedge is set so as to throw q parts of the stream to the left and p parts to the right (of the observer). The

wedges 2 and 3 are set so as to divide the resultant streams in the same proportions. Thus wedge 2 throws q^2 parts of the original material to the left and qp to the right, wedge 3 throws pq parts of the original material to the left and p^2 to the right. The streams passing these wedges are therefore in the ratio of $q^2 : 2qp : p^2$. The next row of wedges is again set so as to divide these streams in the same proportions as before, and the four streams that result will bear the proportions $q^3 : 3q^2p : 3qp^2 : p^3$. The final set, at the heads of the vertical strips, will give the streams proportions $q^4 : 4q^3p : 6q^2p^2 : 4qp^3 : p^4$, and these streams will accumulate between the strips and give a representation of the binomial by a kind of histogram, as shown. Of course as many rows of wedges may be provided as may be desired.

FIG. 10.2.—The Pearson-Galton Binomial Apparatus.

This kind of apparatus was originally devised by Sir Francis Galton (ref. (170)) in a form that gave roughly the symmetrical binomial, a stream of shot being allowed to fall through rows of nails, and the resultant streams being collected in partitioned spaces. The apparatus was generalised by Karl Pearson, who used rows of wedges fixed to movable slides, so that they could be adjusted to give any ratio of $q : p$ (ref. (174)).

10.14. It must not be forgotten that although we have spoken in **10.12** of the skewness and kurtosis of the binomial distribution, it is essentially discontinuous. This is a serious limitation.

Consider, for example, the frequency-distribution of the number of male births in batches of 10,000 births, the mean number being, say, 5100. The distribution will be given by the terms of the series $(0.49 + 0.51)^{10,000}$, and the standard deviation is, in round numbers, 50 births. The distribution will therefore extend to some 150 births or more on either side of the mean number, and in order to obtain it we should have to calculate some 300 terms of a binomial series with an exponent of 10,000! This would not only be practically impossible without the use of certain methods of approximation, but it would give the distribution in quite unnecessary detail: as a matter of practice, we should not have compiled a frequency-distribution by single male births, but should certainly have grouped our observations, taking probably 10 births as the class-interval. We want, therefore, to replace the binomial polygon by some continuous curve, having approximately the same ordinates, the curve being such that the area between any two ordinates y_1 and y_2 will give the frequency of observations between the corresponding values of the variable x_1 and x_2.

Limiting Form of the Binomial for Large n.

10.15. When n becomes large, each term of the binomial becomes small. We are, however, concerned with the sum of the terms falling within certain ranges, and these will not be small in general.

Let us consider first of all the case when p and q are equal. The terms of the series are :

$$N(\tfrac{1}{2})^n \left\{ 1 + n + \frac{n(n-1)}{1 \cdot 2} + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} + \ \ldots \right\}$$

The frequency of m successes is

$$N(\tfrac{1}{2})^n \frac{n!}{m! \, (n-m)!}$$

and the frequency of $m+1$ successes is derived from this by multiplying it by $(n-m)/(m+1)$. The latter frequency is therefore greater than the former so long as

$$n - m > m + 1$$

or

$$m < \frac{n-1}{2}$$

Suppose, for simplicity, that n is even, say equal to $2k$; then the frequency of k successes is the greatest, and its value is

$$y_0 = N(\tfrac{1}{2})^{2k} \frac{(2k)!}{k! \, k!} \qquad . \qquad . \qquad . \qquad (10.6)$$

The polygon tails off symmetrically on either side of this greatest ordinate. Consider the frequency of $k + x$ successes ; the value is

$$y_x = N(\tfrac{1}{2})^{2k} \frac{(2k)!}{(k+x)! \, (k-x)!} \qquad . \qquad . \qquad (10.7)$$

and therefore

$$\frac{y_x}{y_0} = \frac{(k)(k-1)(k-2) \ \ldots \ (k-x+1)}{(k+1)(k+2)(k+3) \ \ldots \ (k+x)}$$

$$= \frac{\left(1 - \dfrac{1}{k}\right)\left(1 - \dfrac{2}{k}\right)\left(1 - \dfrac{3}{k}\right) \ \ldots \ \left(1 - \dfrac{x-1}{k}\right)}{\left(1 + \dfrac{1}{k}\right)\left(1 + \dfrac{2}{k}\right)\left(1 + \dfrac{3}{k}\right) \ \ldots \ \left(1 + \dfrac{x-1}{k}\right)\left(1 + \dfrac{x}{k}\right)} \qquad . \qquad (10.8)$$

Now let us approximate by assuming that k is very large, and indeed large compared with x, so that $(x/k)^2$ may be neglected compared with (x/k). This assumption does not involve any difficulty, for we need not consider values of x much greater than three times the standard deviation or $3\sqrt{k/2}$, and the ratio of this to k is $3/\sqrt{2k}$, which is necessarily small if k be large. On this assumption we may apply the logarithmic series

$$\log_e (1 + \delta) = \delta - \frac{\delta^2}{2} + \frac{\delta^3}{3} - \frac{\delta^4}{4} + \ \ldots$$

12

to every bracket in the fraction (10.8), and neglect all terms beyond the first.   To this degree of approximation,

$$\log_e \frac{y_x}{y_0} = -\frac{2}{k}(1 + 2 + 3 + \ldots + \overline{x-1}) - \frac{x}{k}$$

$$= -\frac{x(x-1)}{k} - \frac{x}{k}$$

$$= -\frac{x^2}{k}$$

Therefore, finally

$$y_x = y_0 e^{-\frac{x^2}{k}} = y_0 e^{-\frac{x^2}{2\sigma^2}} \qquad . \qquad . \qquad . \qquad (10.9)$$

where, in the last expression, the constant k has been replaced by the standard deviation σ, for $\sigma^2 = k/2$.

10.16.   The case when p is not equal to q may be treated in a somewhat similar way but is slightly more complicated.

As before, the frequency of m successes is

$$N \times {}^nC_m q^{n-m} p^m$$

$$= N \frac{n!}{m!\,(n-m)!} q^{n-m} p^m$$

The frequency of $(m+1)$ successes is derived by multiplying this expression by $\dfrac{n-m}{m+1} \cdot \dfrac{p}{q}$, and hence is greater than the former if

$$\frac{n-m}{m+1} \cdot \frac{p}{q} > 1$$

or

$$m < np - q$$

Let us assume that np is a whole number.   Since n is going to tend to infinity, this really imposes no limitation on our work.

The maximum frequency is, then,

$$y_0 = N \frac{n!}{(np)!\,(nq)!} q^{nq} p^{np} \qquad . \qquad . \qquad (10.10)$$

The frequency of $pn + x$ successes is

$$y_x = N \frac{n!}{(np+x)!\,(nq-x)!} q^{nq-x} p^{np+x} \qquad . \qquad . \qquad (10.11)$$

Hence,

$$\frac{y_x}{y_0} = \frac{np!\,nq!}{(np+x)!\,(nq-x)!} q^{-x} p^x \qquad . \qquad . \qquad (10.12)$$

Now, by an important theorem due to James Stirling (1730), if n be large, we have approximately

$$n! = \sqrt{2n\pi}\, n^n e^{-n}$$

Applying this formula here:

$$\frac{y_x}{y_0} = \frac{\sqrt{2np\pi}(np)^{np}e^{-np}\sqrt{2nq\pi}(nq)^{nq}e^{-nq}p^x}{\sqrt{2(np+x)\pi}(np+x)^{np+x}e^{-np-x}\sqrt{2(nq-x)\pi}(nq-x)^{nq-x}e^{-nq+x}q^x}$$

which reduces to

$$\frac{y_x}{y_0} = \frac{1}{\left(1+\dfrac{x}{np}\right)^{np+x+\frac{1}{2}}\left(1-\dfrac{x}{nq}\right)^{nq-x+\frac{1}{2}}}$$

Hence,

$$\log_e\left(\frac{y_x}{y_0}\right) = -(np+x+\tfrac{1}{2})\log_e\left(1+\frac{x}{np}\right)$$

$$-(nq-x+\tfrac{1}{2})\log_e\left(1-\frac{x}{nq}\right)$$

$$= -\left(np+x+\tfrac{1}{2}\right)\left(\frac{x}{np}-\frac{x^2}{2n^2p^2}+\frac{x^3}{3n^3p^3}+\cdots\right)$$

$$-\left(nq-x+\tfrac{1}{2}\right)\left(-\frac{x}{nq}-\frac{x^2}{2n^2q^2}-\frac{x^3}{3n^3q^3}-\cdots\right)$$

After a little rearrangement this becomes:

$$\log_e\left(\frac{y_x}{y_0}\right) = -\frac{x^2}{2npq}+\frac{x^2(p^2+q^2)}{4n^2p^2q^2}+\frac{p-q}{2npq}x+\frac{q^2-p^2}{6n^2p^2q^2}x^3$$

$$+ \text{ terms of order } \frac{1}{n^3} \text{ and higher}$$

Since $q+p=1$, we have, neglecting the terms of order $\frac{1}{n^3}$ and higher, which are small compared with the others when n is large:

$$\log_e\left(\frac{y_x}{y_0}\right) = -\frac{x^2}{2npq}+\frac{x^2(p^2+q^2)}{4n^2p^2q^2}+\frac{q-p}{2npq}\left(-x+\frac{x^3}{3npq}\right) \quad . \quad (10.13)$$

Put, as before, $npq=\sigma^2$, where σ is the standard deviation of the binomial. If n be large, the second term is small compared with the first. Further, since we need not consider values of $\frac{x}{\sigma}$ much greater than 3, if $\frac{q-p}{\sqrt{npq}}$ be small, we can neglect the whole of the third term. On these assumptions we have:

$$\log_e\frac{y_x}{y_0} = -\frac{x^2}{2\sigma^2}$$

or

$$y_x = y_0 e^{-\frac{x^2}{2\sigma^2}} \quad . \qquad . \qquad . \qquad . \quad (10.14)$$

as before.

The expression $\dfrac{q-p}{\sqrt{npq}}$ is merely $\sqrt{\beta_1}$, and so we have in effect simply assumed β_1 small; however much p and q differ we can always make $\sqrt{\beta_1}$ as small as we please by increasing n sufficiently.

10.17. Hence, whether or not p is equal to q, the binomial distribution tends to the form of the continuous curve ((10.9) and (10.14)) when n becomes large, at least for the material part of the range. As a matter of fact, the correspondence between the binomial and the curve is surprisingly close even for comparatively low values of n, provided that p and q are fairly near equality. The student may care to draw the curve with the aid of the tables given at the end of this book (see below, **10.26**) and compare it with some of the simpler binomials drawn to the same scale.

10.18. The curve

$$y = y_0 e^{-\frac{x^2}{2\sigma^2}}$$

is called the **normal curve**. A universe classified according to a continuous variate whose ideal frequency-distribution is a normal curve is called a **normal universe**.

The applications of the normal curve are by no means limited to distributions of the binomial type. Before we refer to its many practical and theoretical applications, however, we shall give a short account of its main properties.

Properties of the Normal Curve.

10.19. The normal curve is obviously symmetrical about the point $x=0$, for its equation is independent of the sign of x. At this point the ordinate has its maximum value. The mean, the median and the mode coincide, and the curve is, in fact, that drawn in fig. 6.5, page 93, and taken as the ideal form of the symmetrical curve.

10.20. The curve is specified completely by defining the mean (the origin of x), the standard deviation σ and the value y_0.

In actual practice, as, for example, when we are trying to fit a normal curve to given data, we are not given y_0 itself, but have to calculate it from the fact that the area of the curve must be equal, on the chosen scale, to the total number of observations. For this reason we wish to find the area under the curve

$$y = y_0 e^{-\frac{x^2}{2\sigma^2}}$$

10.21. From 6.14 it will be seen that the area of a histogram, that is to say, the total number of observations which it represents, is given by

$$\text{Area} = \underset{r=1}{\overset{r=n}{S}} (f_r) \times h$$

where h is the width of the interval, f_r is the frequency in the rth interval and there are n intervals.

As the histogram tends towards the continuous curve the width of the intervals becomes smaller and the number of terms in the summation becomes larger. For the normal curve, which extends to infinity on either side of the mean, the limit to which the sum tends as the intervals

become indefinitely small and the number of terms indefinitely large is
written

$$\int_{-\infty}^{\infty} y_0 e^{-\frac{x^2}{2\sigma^2}} dx$$

the sign $\int$ being a conventional form of the summation sign S and dx
representing the infinitesimally small value of h.

This is the notation of the integral calculus, and the quantity $\int_{-a}^{b} F(x) dx$
is said to be the integral of $F(x)$ with respect to x between the limits $-a$
and $+b$. In this book we shall not use the methods of the integral calculus,
and accordingly it will be necessary for us to state certain results without
proof. It will be sufficient if the student bears in mind that the process of
integration is one of proceeding to the limit in cases of straightforward
summation with which he is already familiar.

10.22. The area of the curve

$$y = y_0 e^{-\frac{x^2}{2\sigma^2}}$$

is then

$$\int_{-\infty}^{\infty} y_0 e^{-\frac{x^2}{2\sigma^2}} dx$$

and this is equal to

$$y_0 \sigma \times \sqrt{2\pi} = 2 \cdot 506627 y_0 \sigma$$

Hence the curve

$$y = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

has unit area, and for this reason the equation of the normal curve is usually
written in the standard form

$$y = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \qquad . \qquad . \qquad . \qquad . \qquad (10.15)$$

From this the form corresponding to a distribution of any given frequency
is immediately written down. In fact, if the frequency is N, the corre-
sponding normal curve is

$$y = \frac{N}{\sigma \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \qquad . \qquad . \qquad . \qquad . \qquad (10.16)$$

Constants of the Normal Curve.

10.23. The mean of the curve is, as we have seen, located at the origin.
If we wish to write the curve with reference to some other point as origin,
we can do so in the form

$$y = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-m)^2} \qquad . \qquad . \qquad . \qquad (10.17)$$

where m is the excess of the mean over the value chosen as origin.

The standard deviation of the curve is σ, and the variance is accordingly σ^2.

The higher moments are calculated by the processes of the integral calculus. Since the nth moment about the mean is given by

$$\mu_n = S(fx^n)$$

we have, proceeding to the limit, that the nth moment of the normal curve is

$$\mu_n = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^n e^{-\frac{x^2}{2\sigma^2}} dx$$

If n is odd this vanishes, as it must for any symmetrical curve. If n is even we have:

$$\mu_n = \frac{n!}{2^{\frac{n}{2}}\left(\frac{n}{2}\right)!} \sigma^n \qquad . \qquad . \qquad . \qquad . \qquad (10.18)$$

and hence,

$$\mu_4 = \frac{4 \cdot 3 \cdot 2}{2 \cdot 2 \cdot 2} \sigma^4 = 3\sigma^4 \qquad . \qquad . \qquad . \qquad (10.19)$$

10.24. From these results it follows that

$$\left.\begin{array}{l} \beta_1 = \gamma_1 = 0 \\ \beta_2 = 3, \quad \gamma_2 = 0 \end{array}\right\} \qquad . \qquad . \qquad . \qquad (10.20)$$

i.e. the normal curve has zero kurtosis. This is, in fact, the origin of the choice of the apparently arbitrary value 3 in the definitions of platy- and lepto-kurtosis (**9.14**).

We may also state without proof the important result that all semin-variants of the normal curve of orders higher than the second vanish identically.

10.25. The mean deviation of the normal curve is

$$\sigma\sqrt{\frac{2}{\pi}} = 0 \cdot 79788 \ldots \sigma$$

This is the origin of the rule given in **8.21**, that the mean deviation is approximately $\frac{4}{5}$ of the standard deviation. The result is true of the normal curve, and very approximately true of curves which do not differ markedly from the normal form. The rules that a range of 6 times the standard deviation includes the great majority of the observations (**8.12**) and that the quartile deviation is about $\frac{2}{3}$ of the standard deviation (**8.24**) were also suggested by the properties of the normal curve (see below, **10.28** and **10.29**).

Ordinates of the Normal Curve.

10.26. The normal curve is so important that tables have been prepared to give (1) the ordinate of the curve corresponding to any given value of x, *i.e.* the values of $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, and (2) the areas of the curve to the

right and the left of any given ordinate, *i.e.* the values of $\dfrac{1}{\sqrt{2\pi}}\displaystyle\int_x^\infty e^{-\frac{x^2}{2}}dx$

and $\dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{\infty}^{x} e^{-\frac{x^2}{2}}dx$. Table 1 of the Appendix gives the values of the ordinate for values of x proceeding by steps of one-tenth of the standard deviation. The values are, of course, the same for positive as for negative values of x. More extended tables will be found in "*Tables for Statisticians and Biometricians, Part I.*"

The ordinate of any normal curve corresponding to a specified value of the variate is easily obtained from the table, as may be seen from the following example :—

Example 10.3.—To find the ordinate of the normal curve given by

$$y = \frac{10,000}{4\sqrt{2\pi}}e^{-\frac{x^2}{32}}$$

corresponding to the variate value $x = 7$.

Here

$$N = 10,000, \qquad \sigma = 4$$

Altering the value of σ is equivalent to altering the scale of x. The ordinate in this curve corresponding to $x = 7$ will be the same as the ordinate of the curve of unit s.d. corresponding to $x = \frac{7}{4} = 1\cdot75$.

From Appendix Table 1, when

$$x = 1\cdot8 \qquad y = 0\cdot07895$$
$$x = 1\cdot7 \qquad y = 0\cdot09405$$

Hence, by simple interpolation, when

$$x = 1\cdot75 \qquad y = 0\cdot08650$$

The ordinate is 10,000 times this

$$= 865$$

The true value, to the nearest unit, obtained by interpolation to second differences, or direct from more extended tables, is 863.

Area of the Normal Curve—the Probability Integral.

10.27. A table of the areas of the normal curve cut off by ordinates at specified values of x is given in Table 2 of the Appendix. As in the case of the table of ordinates, this table is applicable to all normal curves, whatever the value of their standard deviation, the areas cut off on $y = \dfrac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ by ordinates at x being the same as those cut off on $y = \dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}$ by ordinates at $\dfrac{x}{\sigma}$. More extended tables will again be found in "*Tables for Statisticians and Biometricians, Part I.*"

The area of the normal curve to the left of the ordinate at x or, it may be, between the ordinates at 0 and x—conventions differ—is sometimes termed the **probability integral** or the **error function**. These names

arise from the use of the function in the theory of sampling and the theory of errors respectively.

Example 10.4.—Find the frequency represented by the smaller area of the curve $y = \dfrac{10,000}{4\sqrt{2\pi}} e^{-\frac{x^2}{32}}$ cut off by the ordinate at $x = 7$.

Here

$$\sigma = 4, \qquad \frac{x}{\sigma} = 1{\cdot}75 \qquad .$$

For $\dfrac{x}{\sigma} = 1{\cdot}7$ the greater fraction of area $= 0{\cdot}95543$

For $\dfrac{x}{\sigma} = 1{\cdot}8$,,   ,,   ,, $= 0{\cdot}96407$

Hence, by simple interpolation, for

$$\frac{x}{\sigma} = 1{\cdot}75 \text{ the greater fraction of area} = 0{\cdot}95975$$

Hence the smaller fraction $= 1 - 0{\cdot}95975$
$$= \quad 0{\cdot}04025$$

and multiplying this by 10,000, we have the frequency represented, *i.e.* 402·5.

More exactly, by second differences or more extended tables, the value is 400·6.

Example 10.5.—A hundred coins are thrown a number of times. How often approximately in 10,000 throws may (1) exactly 65 heads, (2) 65 heads or more, be expected ?

The number of heads is given by the terms in

$$10,000(\tfrac{1}{2} + \tfrac{1}{2})^{100} \qquad \qquad .$$

The standard deviation is $\sqrt{0{\cdot}5 \times 0{\cdot}5 \times 100} = 5$, $\dfrac{N}{\sigma} = 2000$, and the exponent is large enough for us to be able to take the distribution as normal.

The mean number of heads is 50, and $65 - 50 = 3\sigma$. The frequency of a deviation of 3σ is given at once by Appendix Table 1 as $2000 \times 0{\cdot}00443 = 8{\cdot}86$, or nearly 9 throws in 10,000. A throw of 65 heads will therefore be expected about 9 times.

The frequency of throws of 65 heads *or more* is given by Appendix Table 2, but a little caution must now be used, owing to the discontinuity of the distribution. A throw of 65 heads is equivalent to a range of 64·5–65·5 on the continuous scale of the normal curve, the division between 64 and 65 coming at 64·5. $64{\cdot}5 - 50 = +2{\cdot}9\sigma$, and a deviation of $+2{\cdot}9\sigma$ or more will only occur, as given by the table, 187 times in 100,000 throws, or, say, 19 times in 10,000.

10.28. From the table of areas we can find approximately the position of the quartiles. In fact, we require the value of $\dfrac{x}{\sigma}$ which will give us 0·75

as the greater fraction of the area. From the table we see that this value must lie between 0·6 and 0·7. Simple interpolation gives

$$\left\{ 0\cdot 6 + 0\cdot 1 \frac{2425}{3229} \right\} = 0\cdot 675$$

and more exact interpolation gives

$$\text{Quartile deviation} = 0\cdot 67448975\sigma \qquad . \qquad . \quad (10.21)$$

This is the origin of the rough rule that the semi-interquartile range is usually about $\frac{2}{3}$ of the standard deviation.

10.29. We also observe from the table that an ordinate 3σ from the mean cuts off an area 0·99865 of the whole. The smaller fraction left is therefore 0·00135 of the whole. Since the curve is symmetrical, it follows that a range of 3σ on each side of the mean will cut off all but twice this, *i.e.* all but 0·00270 of the whole. This again is the origin of the rule that such a range includes the great majority of the observations.

The Normal Distribution as an Error Distribution.

10.30. We have deduced the normal distribution as a limiting form of the binomial distribution when n, the exponent, is large. This, however, is only one of the ways in which the normal curve occurs in statistical literature, and Gauss was led to it by a totally different line of reasoning, viz. by inquiring what law of distribution errors of observation should obey in order to make the arithmetic mean of a set of measurements the most likely value of the " true " magnitude.

10.31. Suppose we take a universe of measurements of some magnitude, and consider the universe of deviations from the true value. Let us further suppose that any deviation is the result of the operation of an indefinitely large number of small causes, each producing a small perturbation. Let us assume that the small perturbations are all equal, and that positive and negative perturbations are equally likely.

Then it may be shown that the distribution of errors x about the true value (taken as zero) is given by the law

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

For, if δ is the amount of the perturbation, and positive and negative perturbations are equally likely, the expected frequency of m positive errors and $n - m$ negative errors in N observations is the term $(\frac{1}{2})^m(\frac{1}{2})^{n-m}$ in $N(\frac{1}{2} + \frac{1}{2})^n$, and the actual error is $m\delta - (n - m)\delta = (2m - n)\delta$. Similarly, the frequency of the actual error $\{2(m + 1) - n\}\delta$ is given by the term in $(\frac{1}{2})^{m+1}(\frac{1}{2})^{n-m-1}$; and so on. Proceeding to the limit, as n becomes large, we get the stated result precisely as for the limiting process of **10.15.**

10.32. In the theory of errors it is more customary to write

$$h^2 = \frac{1}{2\sigma^2}$$

so that the distribution becomes

$$y = \frac{h}{\sqrt{\pi}} e^{-h^2 x^2} \qquad . \qquad . \qquad . \qquad . \quad (10.22)$$

h is called the "precision" (*cf.* **8.16**). As h increases, the normal curve becomes narrower and hence h measures in a sense the closeness of the bulk of observations to the true value.

The Occurrence of Normal Distributions in Nature.

10.33. It was found at an early date that error distributions followed the normal law more or less closely, though it must be admitted not with any great exactitude. The fact that many universes, particularly biometrical universes such as those classified according to height and weight, lie distributed round the mean in a humped curve which is not unlike the normal curve, gave rise in the first half of the nineteenth century to keen interest. Although the term " normal " had not then been applied, there appears to have been a feeling that the curve was the ideal to which most distributions should in some degree attain, and that an explanation was demanded if they did not. The normal curve was, in fact, to the early statisticians what the circle was to the Ptolemaic astronomers.

10.34. Workers during the latter half of the nineteenth century were more careful not to let their theories outrun their facts, and as the data accumulated it became evident that the normal distribution was no more usual than any other type. In fact, rather the reverse, so that the occurrence of a normal distribution was to be regarded as something abnormal. " The reader may well ask," says Karl Pearson (ref. (502)), " is it not possible to find material which obeys within probable limits the normal law ? I reply, yes, but this law is not a universal law of nature. We must hunt for cases."

The belief in the validity of the normal law in the theory of errors died harder. " As M. Lippmann once said to me," says Poincaré, in his "*Calcul des Probabilités*," " Everybody believes in the law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact."

10.35. One must, however, be careful not to go too far in seeking to avoid an over-emphasis on the practical occurrence of the normal curve. A certain number of distributions, more particularly those relating to measurements on plants and animals, are approximately of the normal form. As an example, we may take the distribution of Table 6.7, which we show in fig. 10.3 fitted with a normal curve.

Place of the Normal Curve in Theory.

10.36. Strangely enough, the realisation that the normal distribution did not correspond to any widespread natural effect did not diminish its importance in statistical theory. On the contrary, the normal distribution has increased in importance in recent years. It is instructive to consider why this is so.

In the first place, the normal curve and the normal integral have numerous mathematical properties which make them attractive and comparatively easy to manipulate. We have, for instance, already seen that the moments and seminvariants of the normal curve are expressible in simple forms.

Now the normal form is reasonably close to many distributions of the humped type. If, therefore, we are ignorant of the exact nature of a humped distribution, or know the form but find it mathematically intract-

able, we may assume as a first approximation that the distribution is normal and see where this assumption leads us. It is not infrequently found that a universe represented in this way is sufficiently accurately specified for the purposes of the inquiry.

10.37. Secondly, we shall find, when we come to consider sampling distributions, that many of the universes which occur are of the normal form, either exactly or to a satisfactory degree of approximation.

10.38. Thirdly, the theory of the normal curve has been applied to the graduation of curves which are not normal. The Scandinavian school,
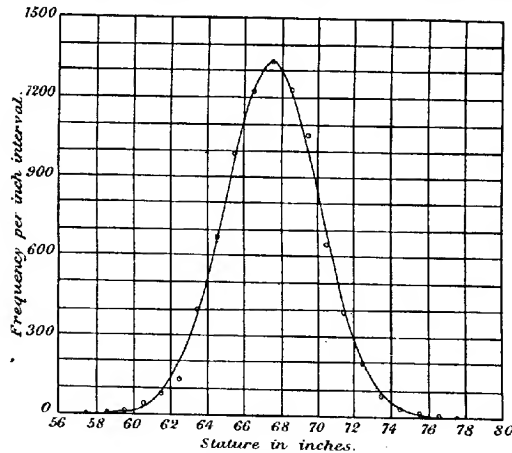
FIG. 10.3.—The Distribution of Stature for Adult Males in the British Isles (fig. 6.6, p. 95), fitted with a Normal Curve. To avoid confusing the figure, the frequency-polygon has not been drawn in, the tops of the ordinates being shown by small circles.

whose interests are mainly actuarial, have developed a technique for expressing a given distribution in the form of an infinite series whose terms depend on the quantity $e^{-\frac{x^2}{2}}$ and certain dependent functions.

10.39. Fourthly, distributions which are not normal can sometimes be brought to a form approximating to the normal by a transformation of the variate. A universe which is skew with respect to a variate x, for instance, might be normal when we take $\sqrt{x}$ as the variate. We gave an example of this kind of effect in Exercise 6.6, page 110, where we saw that a universe of men classified according to their weight was skew, whereas a universe classified according to height (which we may take to be roughly proportional to the cube root of the weight) is nearly normal.

The Poisson Distribution.

10.40. We have found that the limit to the binomial would be a normal curve even if p and q were unequal, provided that n were increased sufficiently to make $(q - p)$ small compared with $\sqrt{npq}$. We now propose

to find the limit to the same series if one of the chances, say q, becomes indefinitely small and n is increased sufficiently to keep nq finite, but not necessarily large—practical values are in fact usually small.

Let us suppose that q is very small and that qn is equal to the finite number m.

In the binomial $(q+p)^n$, the term

$$\frac{n!}{r!\,(n-r)!}q^r p^{n-r}$$
$$=\frac{n!}{r!\,(n-r)!}\left(\frac{m}{n}\right)^r\left(1-\frac{m}{n}\right)^{n-r}$$
$$=\frac{m^r}{r!}\left(1-\frac{m}{n}\right)^n \times \frac{n!}{(n-r)!\,n^r\left(1-\frac{m}{n}\right)^r} \qquad . \qquad . \quad (10.23)$$

Now the limit of $\left(1-\frac{m}{n}\right)^n$ as n becomes large $=e^{-m}$.

Applying Stirling's approximation (10.16) when n is large, the term

$$\frac{n!}{(n-r)!\,n^r\left(1-\frac{m}{n}\right)^r} \qquad . \qquad . \qquad . \quad (10.24)$$

$$=\frac{\sqrt{2\pi n}\,e^{-n}n^n}{\sqrt{2\pi(n-r)}\,e^{-n+r}(n-r)^{n-r}n^r\left(1-\frac{m}{n}\right)^r}$$

$$=\frac{e^{-r}}{\left(1-\frac{r}{n}\right)^n}\cdot\frac{\left(1-\frac{r}{n}\right)^{r-\frac{1}{2}}}{\left(1-\frac{m}{n}\right)^r}$$

Now the limit of $\left(1-\frac{r}{n}\right)^n=e^{-r}$, as we need not consider terms in which r exceeds quantities of the order $\sqrt{nq}$, and the limits of $\left(1-\frac{r}{n}\right)^{r-\frac{1}{2}},\left(1-\frac{m}{n}\right)^r$ are both unity. Hence the limit of (10.24) is unity, and the limit of (10.23) is

$$\frac{m^r e^{-m}}{r!}$$

10.41. Hence the successive terms in the binomial are

$$e^{-m}, \qquad e^{-m}m, \qquad e^{-m}\frac{m^2}{2\,!}, \qquad e^{-m}\frac{m^3}{3\,!}, \text{ etc.}$$

and the limit of $(q+p)^n$ is

$$e^{-m}\left(1+m+\frac{m^2}{2!}+\frac{m^3}{3!}+\ \ldots\right) \qquad . \qquad . \qquad . \quad (10.25)$$

This expression is called **Poisson's distribution**, or Poisson's exponential limit. It was first published by Poisson in 1837, but has subsequently been rediscovered by numerous writers.

Constants of the Poisson Distribution.

10.42. Taking an origin located at the first term of the distribution, we have :

$$\mu_1' = e^{-m}\left[0 + m + \left(\frac{m^2}{2!} \times 2\right) + \left(\frac{m^3}{3!} \times 3\right) + \ldots\right]$$

$$= me^{-m}\left(1 + \frac{m}{1!} + \frac{m^2}{2!} + \ldots\right)$$

$$= me^{-m}e^{m}$$

$$= m$$

$$\mu_2' = e^{-m}\left[0 + m + \left(\frac{m^2}{2!} \times 2^2\right) + \left(\frac{m^3}{3!} \times 3^2\right) + \ldots\right]$$

$$= e^{-m}\left[m + \left(\frac{m^2}{1!} \times 2\right) + \left(\frac{m^3}{2!} \times 3\right) + \ldots\right]$$

$$= me^{-m}\left(1 + \frac{m}{1!}(1+1) + \frac{m^2}{2!}(2+1) + \ldots\right)$$

$$= me^{-m}\left(1 + \frac{m}{1!} + \frac{m^2}{2!} + \ldots + m + \frac{m^2}{1!} + \ldots\right)$$

$$= me^{-m}(e^{m} + me^{m})$$

$$= m(m+1)$$

It may also be shown that

$$\mu_3' = m(m^2 + 3m + 1) = m\{(m+1)^2 + m\}$$
$$\mu_4' = m(m^3 + 6m^2 + 7m + 1)$$

From these results we have immediately :

$$\text{Mean} = m \qquad . \qquad . \qquad . \qquad . \quad (10.26)$$

$$\mu_2 = m(m+1) - m^2$$

$$= m$$

$$\sigma = \sqrt{m} \qquad . \qquad . \qquad . \qquad . \quad (10.27)$$

Hence,

$$\sigma^2 = m = \text{mean}$$

10.43. The third and fourth moments about the mean will be found to be

$$\mu_3 = m \qquad . \qquad . \qquad . \qquad . \quad (10.28)$$
$$\mu_4 = 3m^2 + m \qquad . \qquad . \qquad . \qquad . \quad (10.29)$$

so that

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{m^2}{m^3} = \frac{1}{m} \qquad . \qquad . \qquad . \quad (10.30)$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3m^2 + m}{m^2} = 3 + \frac{1}{m} \qquad . \qquad . \quad (10.31)$$

These results should be compared with the expressions

$$\beta_1 = \frac{(p-q)^2}{npq}$$

$$\beta_2 = 3 + \frac{1-6pq}{pqn}$$

for the binomial. They are, as might be expected, the limits of those expressions when $q = \dfrac{m}{n}$ and n is large.

10.44. We may state without proof that *all* the seminvariants of the Poisson distribution are equal to m.

10.45. Tables of the limit $e^{-m}\dfrac{m^r}{r!}$ for various values of m and r have been published by several authorities. One such set will be found in "*Tables for Statisticians and Biometricians, Part I.*"
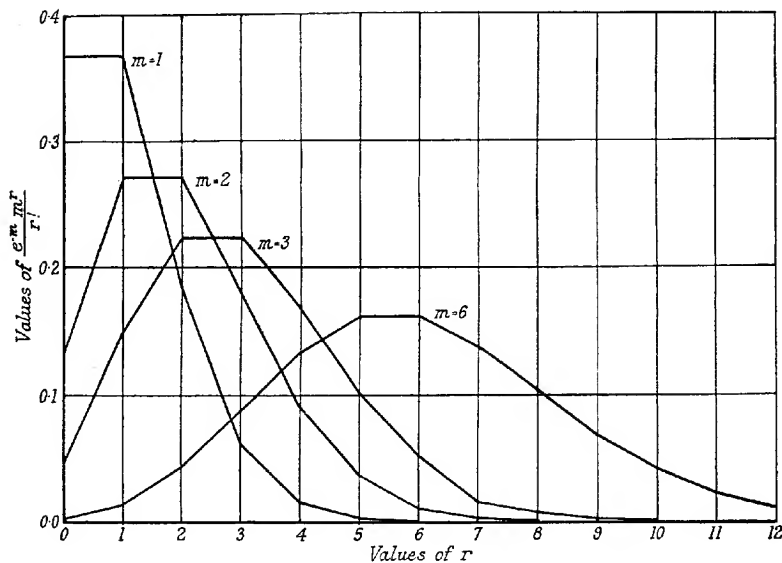
FIG. 10.4.—Frequency-polygons of the Poisson Series for Various Values of m.

The form of the frequency-polygon of the distribution (which, like the binomial and unlike the normal, is discontinuous) can be judged from fig. 10.4, in which the polygons for various values of m are drawn. It will be seen that for low values of m the polygon is very skew, but that for larger values it tends towards a symmetrical form.

10.46. The condition that p or q shall be small, np or nq remaining finite, implies that in practice we should expect to find a Poisson distribution in cases where the chance of any individual being a " success " was small. Such a case might arise, for example, in considering the deaths from a rare disease in a population, the chance of any individual dying from it being small.

10.47. Attention to the fact that comparatively rare events are not haphazard was first directed by Quetelet and von Bortkiewicz. The latter's data of the number of men killed by the kick of a horse in certain Prussian army corps in twenty years (1875–94) have become classical.

The frequency-distribution of the number of deaths in 10 corps per army corps per annum over twenty years was :

Deaths.	Frequency.
0	109
1	65
2	22
3	3
4	1

Here the total number of deaths was 122, and hence the mean deaths per army corps per annum is 0·61. Taking this as m, we find the following values for various numbers of deaths per annum :—

Deaths.	Frequency assigned by Poisson's Limit.
0	108·7
1	66·3
2	20·2
3	4·1
4	0·7 (4 and over)

If we calculate σ^2 for the actual distribution, we find :

$$\sigma = 0·78, \qquad \sigma^2 = 0·6079$$

Hence, σ^2 is nearly equal to the mean, which is in accordance with theory. The agreement is, in fact, very much closer than is usual. Many distributions are now available for the frequency of individuals who have met with 0, 1, 2, . . . accidents, *e.g.* in factories, during a given period of time, and more often than not such distributions give a value of the variance exceeding the mean. This state of affairs can be accounted for on the assumption that the individuals at risk have varying degrees of " accident-proneness," and the assumption has been corroborated by finding that those individuals who have the largest number of accidents in one period are, on the whole, those who have most accidents during a succeeding period.

Another example of the Poisson distribution is given in Exercise 10.17 at the end of this chapter. The early instances of the distribution were nearly all demographic, and for some time it remained more of a curiosity than a useful tool. In 1907, however, " Student " drew attention to a class of hæmacytometer counts to which the distribution seemed appropriate, and since that time it has found several important biological applications. It also appears in problems of controlling road and telephone traffic.

Pearson Curves.

10.48. The process of obtaining the normal curve as a limit of the binomial suggested to Karl Pearson an investigation into a series of analogous curves which may be regarded as limits to skew binomials or to distributions from a finite universe, *e.g.* by drawing r balls at a time from a bag which contains a finite number N of black and white balls in given proportions. One such curve was of the form

$$y = y_0 \left(1 + \frac{x}{a}\right)^{\gamma a} e^{-\gamma x}$$

This set of curves, divided into twelve types, which were later regarded from rather a different standpoint, can be made to fit a large number of the distributions occurring in practice.

In the curve given above, γ, a and the origin can all be obtained from the first three moments. For the other curves of Pearson's system, except some degenerate types, the first four moments are necessary to specify the constants of the curve completely. The distributions considered hitherto have required in addition to the area (number of observations), either the mean only (Poisson) or the mean and standard deviation (normal curve) to determine their constants ; but the principle of fitting for the more general curves remains the same. The actual moments of the curves are equated to the moments expressed in terms of the constants, such as γ and a, which are to be found. For full details of these curves, the method of determining the type to choose and the method of fitting, the student is referred to Elderton's book (ref. (160)).

SUMMARY.

1. If the chance of the success of an event is p, and of its failure q, then, provided that the chance remains constant throughout the trials, the expected frequencies of 0, 1, 2, . . . successes in N sets of n trials are the 1st, 2nd, etc. terms in the binomial

$$N(q + p)^n$$

2. The mean of the binomial is pn and its standard deviation is $\sqrt{npq}$.

3. For the binomial :

$$\beta_1 = \frac{(q - p)^2}{npq}, \qquad \beta_2 = 3 + \frac{1 - 6pq}{pqn}$$

4. If neither p nor q is small, the binomial tends for large values of n to the form

$$y = y_0 e^{-\frac{x^2}{2\sigma^2}}$$

5. This curve, which may also be written

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

is called the normal curve.

6. The standard deviation of the normal curve is σ. Its third moment is zero, and the fourth moment is $3\sigma^4$. Hence,

$$\beta_1 = 0, \qquad \beta_2 = 3$$

All seminvariants higher than the second are zero.

7. In the theory of errors the normal universe is usually written :

$$y = \frac{h}{\sqrt{\pi}} e^{-h^2 x^2}$$

$h = \dfrac{1}{\sigma\sqrt{2}}$ being called the precision.

8. The mean deviation of the normal curve is

$$\sigma\sqrt{\frac{2}{\pi}} = 0\cdot79788 \ \ldots \ \sigma$$

and the quartile deviation (semi-interquartile range) is $0\cdot67448975 \ldots \sigma$.

9. A range 3σ on each side of the mean of the normal curve contains $0\cdot9973$ of the distribution.

10. If p or q is small and one of pn, qn is finite and equal to m, the binomial distribution tends to the limit

$$e^{-m}\left(1 + m + \frac{m^2}{2!} + \ \ldots \ + \frac{m^r}{r!} + \ \ldots \right)$$

This is called the Poisson distribution.

11. The mean of the Poisson distribution is m, and σ^2 also equals m.

12. For the Poisson distribution :

$$\beta_1 = \frac{1}{m}, \qquad \beta_2 = 3 + \frac{1}{m}$$

and all the seminvariants are equal to m.

EXERCISES.

10.1. A perfect cubic die is thrown a large number of times in sets of 8. The occurrence of a 5 or a 6 is called a success. In what proportion of the sets would you expect 3 successes?

10.2. The following data, due to W. F. R. Weldon, show the results of throwing 12 dice 4096 times, a throw of 4, 5 or 6 being called a success :—

Successes.	Frequency.	Successes.	Frequency.
0	—	7	847
1	7	8	536
2	60	9	257
3	198	10	71
4	430	11	11
5	731	12	—
6	948	Total	4096

13

Find the expected frequencies, and compare the actual mean and standard deviation with those of the expected distribution.

10.3. In the previous example find the equation of the normal curve which has the same mean, standard deviation and total frequency as the observed distribution.

Find the frequencies to be expected if the distribution were represented exactly by the ordinates of this curve and compare them with the actual frequencies.

10.4. Assuming that half the population are consumers of chocolate, so that the chance of an individual being a consumer is $\frac{1}{2}$, and assuming that 100 investigators each take ten individuals to see whether they are consumers, how many investigators would you expect to report that three people or less were consumers?

10.5. An irregular six-faced die is thrown, and the expectation that in 10 throws it will give five even numbers is twice the expectation that it will give four even numbers. How many times in 10,000 sets of 10 throws would you expect it to give no even numbers?

10.6. If two normal universes have the same total frequency but the σ of one is k times that of the other, show that the maximum frequency of the first is $\dfrac{1}{k}$ that of the other.

10.7. Find graphically or otherwise the point of inflection of the normal curve, and show that it occurs at a distance σ from the mean ordinate.

10.8. Show that if np be a whole number, the mean of the binomial coincides with the greatest term.

10.9. Show that if two symmetrical binomial distributions of degree n (and of the same number of observations) are so superposed that the rth term of the one coincides with the $(r+1)$th term of the other, the distribution formed by adding superposed terms is a symmetrical binomial of degree $(n+1)$.

[Note.—It follows that if two normal distributions of the same area and standard deviation are superposed so that the difference between the means is small compared with the standard deviation, the compound curve is very nearly normal.]

10.10. Calculate the ordinates of the binomial $1024\,(0\cdot5+0\cdot5)^{10}$, and compare them with those of the normal curve.

10.11. If skulls are classified as *dolichocephalic* when the length-breadth index is under 75, *mesocephalic* when the same index lies between 75 and 80, and *brachycephalic* when the index is over 80, find approximately (assuming that the distribution is normal) the mean and standard deviation of a series in which 58 per cent. are stated to be dolichocephalic, 38 per cent. mesocephalic and 4 per cent. brachycephalic.

10.12. Find the deciles of the normal curve.

10.13. Write down the normal universe which has the same mean and (uncorrected) standard deviation as that of the last column of Table 6.7, page 94, and find the mean deviation and quartile deviation. Compare the results with the corresponding quantities for the actual distribution.

10.14. Proceed similarly for the skew universe of Table 6.8, page 96.

10.15. In Exercise 10.4, if 1000 investigators each choose 100 individuals, how many would you expect to report that more than 60 persons are consumers?

10.16. Taking the universe of screws of Table 6.3, page 84, find the normal universe which has the same standard deviation and a mean of 1 inch. Compare the frequencies given by this universe with the actual frequencies.

10.17. The following data (Lucy Whitaker, ref. (190)) give the number of deaths of women over 85 published in *The Times* during 1910–12 :—

Number of Deaths per day.	Frequency.
0	364
1	376
2	218
3	89
4	33
5	13
6	2
7	1

Find the frequencies of the Poisson distribution which has the same mean as this distribution, and compare your results with the actual frequencies. For the purpose of this example, simple interpolation in the tables given in " *Tables for Statisticians and Biometricians* " is sufficient.

10.18. In the data of the previous exercise calculate the first four semin-variants.

CHAPTER 11.

CORRELATION.

Bivariate Universes.

11.1. In Chapters 6 to 10 we considered the members of a universe classified according to the values of a single variable; and we saw how they could be grouped into a frequency-distribution whose characteristics could be described by certain constants. We have now to proceed to the case of two variables, in which each member of the universe will exhibit two values, one for each of the variables under consideration.

A universe of this kind is called a **bivariate universe.** One of our main topics will be the way in which the two variables are related in the universe.

11.2. If the corresponding values of the two variables are noted for each member, the methods of classification employed in the previous chapters may be applied to both variables. We can thus group our data into a table of double entry, or contingency table (Chapter 5), showing the frequencies of pairs of values lying within given class-intervals. Six such tables are given below as illustrations for the following variables : Table 11.1, two measurements on a shell; Table 11.2, ages of husbands and their wives in marriages taking place in England and Wales in 1933 ; Table 11.3, statures of fathers and their sons ; Table 11.4, age and yield of milk in cows ; Table 11.5, the rate of discount and ratio of reserves to deposits in American banks ; Table 11.6, the proportion of male to total births and the total numbers of births in the registration districts of England and Wales.

Arrays and Correlation Tables.

11.3. Each row in such a table gives the frequency-distribution of the first variable for the members of the universe in which the second variable lies within the limits stated on the left of the row. Similarly for the columns. As " columns " and " rows " are distinguished only by the accidental circumstances of the one set running vertically and the other horizontally, and the difference has no statistical significance, the word **array** has been suggested as a convenient term to denote either a row or a column.

If the values of X in one array are associated with values of Y in an interval centred at Y_n, then Y_n is called the **type** of the array.

11.4. A grouped frequency-distribution of the type of Tables 11.1 to 11.6 may then be termed a bivariate frequency-distribution ; but if we are particularly interested in the relationship between the two variates it is sometimes called a **correlation table.** The difference between a correlation table and a contingency table lies in the fact that the latter term may

be, and usually is, applied to tables classified according to unmeasured quantities or imperfectly defined intervals.

TABLE 11.1.—*Correlation between* (1) *Antero-posterior and* (2) *Dorso-ventral Diameter in Lower Valve of Pecten opercularis.* (Condensed from a Table given by C. B. Davenport, *Proc. Amer. Ac.*, vol. 39, 1903, p. 149.) Measurements in millimetres.

(1) Antero-posterior diameter, mm. (rows) — (2) Dorso-ventral diameter, mm. (columns)

(1)\(2)	37–39	40–42	43–45	46–48	49–51	52–54	55–57	58–60	61–63	64–66	67–69	70–72	Total
Total	4	19	48	59	44	105	147	80	21	6	2	2	537
76–78											1	2	3
73–75													1
70–72									1	1	1		3
67–69									3	5			8
64–66								7	13				20
61–63							25	59	4				88
58–60						8	90	14					112
55–57					3	68	32						103
52–54				1	17	29							47
49–51				22	22								44
46–48			12	35	2								49
43–45		6	35	1									42
40–42		12	1										13
37–39	4	1											5

11.5. We need add very little to what was said in Chapter 6 about the choice and magnitude of class-intervals and the classification of data. When the intervals have been fixed, the table is readily compiled from the raw material by taking a large sheet of paper ruled with arrays properly

TABLE 11.2.—*Correlation between Ages of* (1) *Husband and* (2) *Wife in Marriages in England and Wales in 1933.* (Figures in hundreds— certain marriages in which no age specified are omitted. Data from Registrar-General's Statistical Review of England and Wales for 1933, Tables, Part II, Civil.)

(1) Age of Husband (Years).

	15–	20–	25–	30–	35–	40–	45–	50–	55–	60–	65–	70–	75–	Total.
15–	33	189	56	8	2	—	—	—	—	—	—	—	—	288
20–	18	682	585	106	19	5	2	1	—	—	—	—	—	1418
25–	1	140	511	179	40	14	6	3	1	1	—	—	—	896
30–	—	11	75	101	42	20	10	5	2	1	1	—	—	268
35–	—	2	10	24	28	19	13	8	5	2	1	—	—	112
40–	—	—	1	5	9	14	12	10	6	4	2	1	—	64
45–	—	—	—	1	2	5	9	9	7	4	3	1	—	42
50–	—	—	—	—	—	1	3	7	6	5	3	1	—	26
55–	—	—	—	—	—	—	1	3	5	4	3	1	—	17
60–	—	—	—	—	—	..	...	1	1	4	3	2	—	11
65–	—	—	—	—	—	—	—	—	1	1	3	2	1	8
70–	—	—	—	—	—	—	—	—	—	—	1	1	1	3
Total	52	1024	1238	424	143	78	56	47	34	26	20	9	2	3153

(2) Age of Wife (Years).

headed in the same way as the final table and entering a small mark in the compartment corresponding to the variate values exhibited by each individual. If facility of checking be of great importance, each pair of recorded values may be entered on a separate card and these dealt into little packs on a board ruled in squares, or into a divided tray ; each pack can then be run through to see that no card has been mis-sorted. The difficulty as to the intermediate observations—values of the variables corresponding to divisions between class-intervals—will be met in the same way as before if the value of one variable alone be intermediate, the unit of frequency being divided between two adjacent compartments. If both values of the pair be intermediates, the observation must be divided between *four* adjacent compartments, and thus quarters as well as halves may occur in the table, as, *e.g.*, in Table 11.3. In this case the statures of fathers and sons were measured to the nearest quarter-inch and subsequently grouped by 1-inch intervals : a pair in which the recorded stature of the father is 60·5 in. and that of the son 62·5 in. is accordingly entered as 0·25 to each of the four compartments under the columns 59·5–60·5, 60·5–61·5, and the rows 61·5–62·5, 62·5–63·5.

Frequency-surface and Stereogram.

11.6. The distribution of frequency for two variables may be represented by a surface in three dimensions in the same way as the frequency-distribution for a single variable may be represented by a curve in two. We may imagine the surface to be obtained by erecting at the centre of every compartment of the correlation table a vertical of length proportionate to the frequency in that compartment, and joining up the tops of the verticals. If the compartments were made smaller and smaller while the class-frequencies remained finite, the irregular figure so obtained would approximate more and more closely towards a continuous curved surface —a **frequency-surface**—corresponding to the frequency-curves for single

variables of Chapter 6. The volume of the frequency-solid over any area drawn on its base gives the frequency of pairs of values falling within that

TABLE 11.3.—*Correlation between* (1) *Stature of Father and* (2) *Stature of Son: 1 or 2 Sons only of each Father.* Measurements in inches. (From Karl Pearson and Alice Lee, *Biometrika*, vol. 2, 1903, p. 415.)

(2) Stature of Son	\ (1) Stature of Father →	74·5–75·5	73·5–74·5	72·5–73·5	71·5–72·5	70·5–71·5	69·5–70·5	68·5–69·5	67·5–68·5	66·5–67·5	65·5–66·5	64·5–65·5	63·5–64·5	62·5–63·5	61·5–62·5	60·5–61·5	59·5–60·5	58·5–59·5	Total.
59·5–60·5														·5					2
60·5–61·5											1			·5					1·5
61·5–62·5						·25			1			1				·25			3·5
62·5–63·5						1·25	·25	·5		·25	·25	2	2·25		·25	·25			20·5
63·5–64·5					2·5	·75	1·5	1·25	·5	5	4	4·25	3	2·25	·25	·25	1		38·5
64·5–65·5					3·25	2·5	3·5	1·25	2·75	9·25	8	9·5	5·25	3·75			2		61·5
65·5–66·5				1	8·5	13·75	5·25	6·5	3	10·75	13·5	9·5	3·5	2·25	1·5	1			89·5
66·5–67·5			1	·5	10	13·25	12·5	16	7·5	16·75	10	13·75	6·25	4·75	·5	·5			148
67·5–68·5			2·25	9·5	14·5	21·5	29·5	19·5	17·5	26·5	19·75	10		2	1	1·5			173·5
68·5–69·5		1·5	2·25	8·5	10·75	20·75	20	23·5	25·75	24·25	16·25	·5			1·5				149·5
69·5–70·5	1		3·5	6·25	10	11·25	22·5	24	31·5	18·25	12·75	2·5			1				128
70·5–71·5	1	·5	5	8·5	7·5	6	14·75	19·5	16	18·75	6·75	3·25	1		1·5				108
71·5–72·5			2·70	6·25	6·5	2·5	10·75	11	11·75	8·75	5	·25			1				63
72·5–73·5	·5	·5	3·25	5·25	2·5		6·5	7·75	10·75	1·25	3								42
73·5–74·5	2		1·70	·75	·5	1	2·25	7·5	2·5	·75	·75								29
74·5–75·5		·5		1		1	·25	5·25		1·5	1·5								8·5
75·5–76·5			1·5				·25	1·25											4
76·5–77·5			·75	·25			1	1·25											4
77·5–78·5			·25	·25				1											·5
78·5–79·5																			
Total	8·5	4	28·5	49	78	116	141·5	164	137·5	142	95·5	61·5	33·5	17	8	3·5	3	**1078**

area, just as the area of the frequency-curve over an interval of the base line gives the frequency of observations within that interval.

11.7. Similarly, a figure analogous to the frequency-polygon or the histogram may be constructed by drawing the frequency-distributions for

all arrays of the one variable, to the same scale, on sheets of cardboard, cutting-out and erecting the cards vertically on a base-board at equal

TABLE 11.4.—*Correlation between* (1) *Age in Years and* (2) *Yield of Milk per Week in 4912 Ayrshire Cows.* (Data from J. F. Tocher, "An Investigation of the Milk Yield of Dairy Cows," *Biometrika*, vol. 20B, 1928, pp. 106–244.)

(1) Age in Years.

Age \ Yield	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	Totals
3				3	2	2	9	11	15	16	11	9	8	3	5	1	2	3										112
4		2	5	10	25	76	76	115	149	148	146	117	97	63	42	19	20	10	7	2								1129
5	2	1	8	17	29	57	79	119	131	132	112	107	93	63	33	33	15	13	7	2	2		2					1047
6			1	7	9	8	38	43	83	74	94	83	113	79	88	49	38	34	22	7	9	1	2		1	2		812
7		1	3	1	5	9	23	34	59	58	73	87	69	70	45	38	27	17	4	5	4	4						636
8			4	2	9	24	23	34	49	51	51	49	32	27	19	20	15	5	5	2	1	2						419
9				1	4	4	7	11	23	32	39	35	25	31	14	17	13	8	2	4	1	3		2				276
10				2	1	6	8	16	15	22	33	30	29	18	17	9	10	4	2	1								223
11		1			2	1	1	4	4	9	12	17	11	11	9	10	12	3	3	2		2	3	2			1	122
12						1	1	1	2	5	7	6	6	10	10	7	3	7	2	4	3							75
13							3	1	4	5	5	2	3	4	1	1	1		1				1					32
14				1		1		2			1	3	3		2	2												15
15								1		1	1	1		1	2													7
16					1													1										2
17								1	1		1	1																4
18												1																1
Totals	1	5	13	33	71	151	236	339	499	552	585	586	496	448	284	214	153	112	58	35	13	15	4	5	2	1	1	4912

(2) Yield of Milk per Week (Gallons). (Central Value of Interval.)

distances apart, or by marking out a base-board in squares corresponding to the compartments of the correlation table, and erecting on each square a rod of wood of height proportionate to the frequency. Such solid repre-

sentations of frequency-distributions for two variables are sometimes termed **stereograms**.

TABLE 11.5.—Correlation between (1) Call Discount Rates and (2) Percentage of Reserves on Deposits in New York Associated Banks (Weekly Returns). (From "Statistical Studies in the New York Money Market," by J. P. Norton. Publications of the Department of the Social Sciences, Yale University; The Macmillan Co., 1902.) Note that, after the column headed 8 per cent, blank columns have been omitted to save space.

(2) Percentage Ratio of Reserves to Deposits	Total	(1) Call Discount Rates																				
		25	20	15	12	10	9	8	7·5	7	6·5	6	5·5	5	4·5	4	3·5	3	2·5	2	1·5	1
21	2						1									1						
22	1													1								
23	1													1								
24	9	1		1		2		3				2										
25	42			1		2	1	6		2	1	11	4	4	6	2	1					
26	85	2	1	1	1	2	2			7	4	11	6	16	12	13	6	2				
27	124	1						1		4	3	9	9	19	17	15	12	14	9	10		
28	115	1				1				3	2	2	1	7	3	7	11	20	23	30	1	
29	109									2				1	3	6	3	16	17	48	5	3
30	53													2		4	4	8	10	12	12	1
31	36													1	1	2	2	4	2	6	10	8
32	53														1			5	8	10	14	15
33	32															2			1	4	8	15
34	14																			1	11	2
35	14																			1	5	8
36	10																			1	2	7
37	9																1					8
38	11																				1	9
39	21																				2	19
40	15																				8	7
41	10																				3	7
42	10																				2	8
43	1																					1
44	1																					1
45	2																					2
Total	780	4	1	3	1	7	4	10	—	18	10	35	20	52	45	52	40	69	70	125	93	121

11.8. It is impossible, however, to group the majority of frequency-surfaces, in the same way as the frequency-curves, under a few simple types : the forms are too varied. The simplest ideal type is one in which every section of the surface is a symmetrical curve—the first type of

Chapter 6, fig. 6.5, page 93. Like the symmetrical distribution for the single variable, this is a very rare form of distribution in economic statistics,

TABLE 11.6.—Showing the Number of Registration Districts in England and Wales exhibiting (1) a Given Proportion of Male Births, (2) a Given Total Number of Births during the Decade 1881–90. (The Data as to Total Births and Numbers of Male and Female Births from Decennial Supplement to Report of the Registrar-General. Table from H. D. Vigor and G. U. Yule, Jour. Roy. Stat. Soc., vol. 69, 1906.)

(1) Proportion of Male Births per 1000 of all Births.

	0–4	4–8	8–12	12–16	16–20	20–24	24–28	28–32	32–36	36–40	40–44	44–48	48–52	52–56	56–60	60–64	64–68	68–72	72–76	76–80	80–84	84–88	88–92	92–96	96–100	100–04	104–08	148–52	Total
Total.	149	204	85	48	28	15	15	12	7	8	8	11	7	7	5	5	6	3	4	1	3	2	1	2		1		1	632
543–45	1																												1
540–42	1																												1
537–39																													1
534–36	2	1																											3
531–33	1																												1
528–30	1																												1
525–27	3		1																										4
522–24	5	4	1	1																									11
519–21	9	10	4	1	1	1																							26
516–18	12	18	7	6	1	1	1			1																			47
513–15	14	36	13	8	4	3	1	1	2	1	1																		88
510–12	51	42	18	12	9	4	8	8			4	3	3		3		3	1			1			2					129
507–09	12	29	17	8	6	5	6	3	4	5	6	5		1	1		3		1	1	2	1				1	1		125
504–06	19	27	15	10	5	1	3	3		2	2	1	3		2	1						1							98
501–03	18	20	6	1		1	2																						46
498–500	9	7	1				1																						18
495–97	8	5	2																										15
492–94	4	2																											6
489–91	2	2																											4
486–88	2																												2
483–85	1	1																											3
480–82	2																												2
477–79																													1
474–76																													1
471–73	1																												1
468–70																													1
465–67	1																												1
Total																													

(2) Total Number of Births in District (000's omitted) during Decade.

but approximate illustrations may be drawn from anthropometry. Fig. 11.1 shows the ideal form of the surface, somewhat truncated, and fig. 11.3 the distribution of Table 11.3, which approximates to the same type— the difference in steepness is, of course, merely a matter of scale. The

maximum frequency occurs in the centre of the whole distribution, and
the surface is symmetrical round the vertical through the maximum, equal
frequencies occurring at equal distances from the mode on opposite sides.

TABLE 11.7.—*Showing the Monthly Index-numbers of Prices of* (1) *Animal Feeding-stuffs
and* (2) *Home-grown Oats in England and Wales for 1931–1935.* The index-numbers
are based on prices in corresponding months of 1911–13. (Data from Agricultural
Market Report for England and Wales.)

Month.	Index of Feeding-stuffs Price.	Index of Oats Price.	Month.	Index of Feeding-stuffs Price.	Index of Oats Price.
1931 Jan.	78	84	1933 July	85	75
Feb.	77	82	Aug.	83	79
Mar.	85	82	Sept.	80	78
Apr.	88	85	Oct.	78	78
May	87	89	Nov.	80	76
June	82	90	Dec.	83	75
July	81	88			
Aug.	77	92	1934 Jan.	82	80
Sept.	76	83	Feb.	83	91
Oct.	83	89	Mar.	85	87
Nov.	97	98	Apr.	83	84
Dec.	93	99	May	82	81
			June	85	83
1932 Jan.	95	102	July	88	83
Feb.	97	102	Aug.	101	92
Mar.	102	105	Sept.	102	98
Apr.	99	105	Oct.	98	94
May	97	107	Nov.	96	94
June	94	107	Dec.	98	95
July	94	101			
Aug.	97	106	1935 Jan.	98	100
Sept.	92	96	Feb.	92	99
Oct.	89	90	Mar.	92	96
Nov.	90	85	Apr.	90	98
Dec.	90	81	May	88	97
			June	86	98
1933 Jan.	92	84	July	83	99
Feb.	91	85	Aug.	80	92
Mar.	90	84	Sept.	81	90
Apr.	86	81	Oct.	86	89
May	85	76	Nov.	83	87
June	85	77	Dec.	82	83

The next simplest type of surface corresponds to the second type of
frequency-curve—the moderately asymmetrical. Most, if not all, of the
distributions of arrays are asymmetrical, and like the distributions of fig. 6.7;
the surface is consequently asymmetrical, and the maximum does not lie
in the centre of the distribution. This form is fairly common, and illustra-
tions might be drawn from a variety of sources—economics, meteorology,
anthropometry, etc. The data of Table 11.4 will serve as an example.
The total distributions and the distributions of the majority of the arrays
are asymmetrical, the rows being markedly so. The maximum frequency
lies towards the upper end of the table in the compartment under the row
headed "16" and column headed "4." The frequency falls off very
rapidly towards the lower ages, and slowly in the direction of old age.

Outside these two forms, it seems impossible to delimit empirically any simple types. Tables 11.5 and 11.6 are given simply as illustrations of two
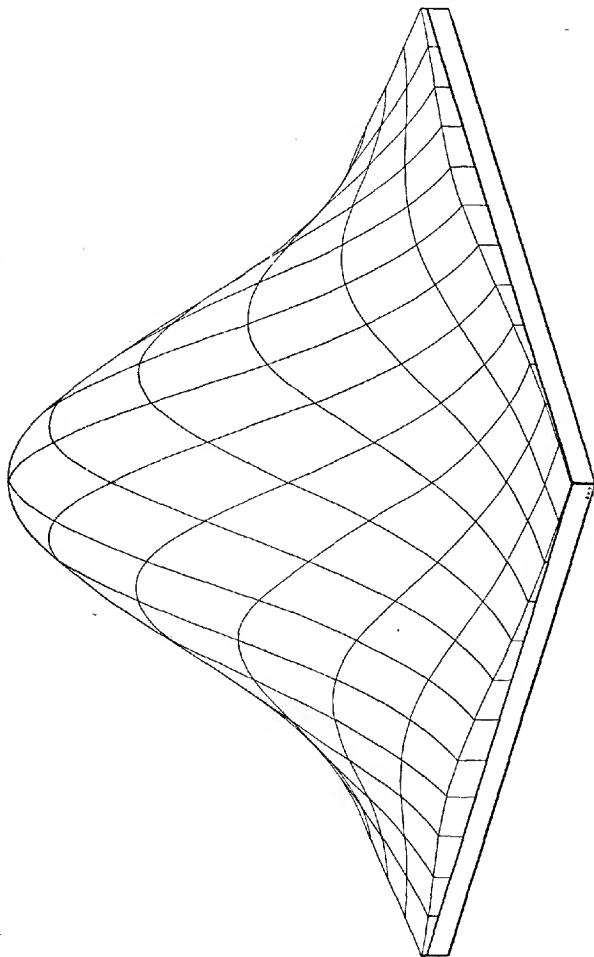
FIG. 11.1.—The Ideal Symmetrical ("Normal") Frequency-surface, with the Extremes Truncated.

very divergent forms. Fig. 11.2 gives a graphical representation of the former by the method corresponding to the histogram of Chapter 6, the frequency in each compartment being represented by a square pillar. The distribution of frequency is very characteristic, and quite different from that of any of the Tables 11.1 to 11.4.
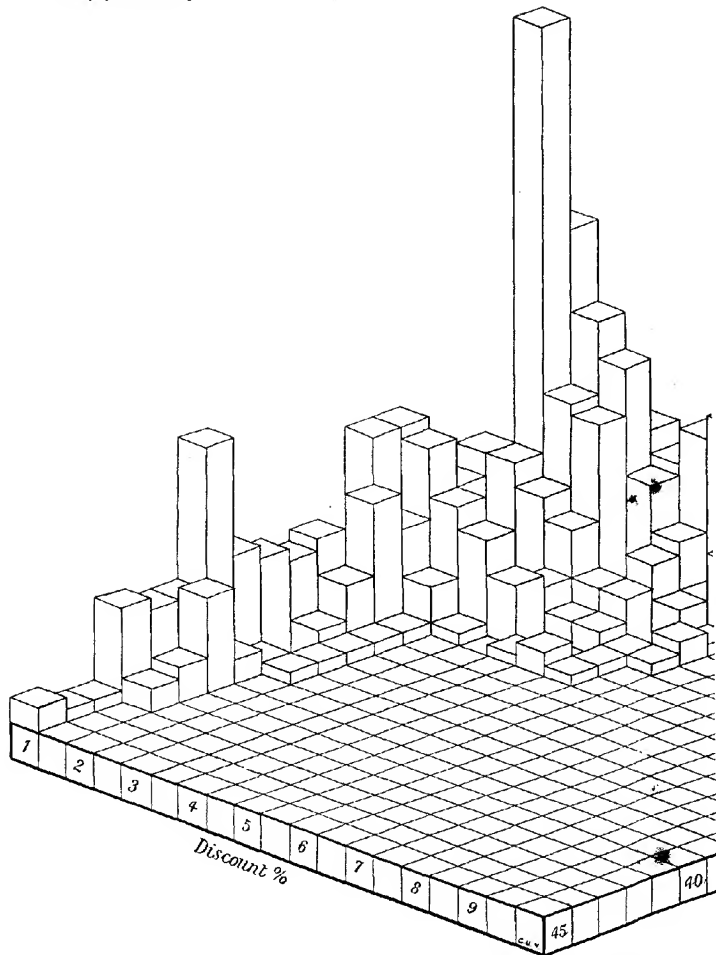
FIG. 11.2—Frequency-surface for the Rate of Discount and Ratio of Reserves to

64

68

72

Son's Stature
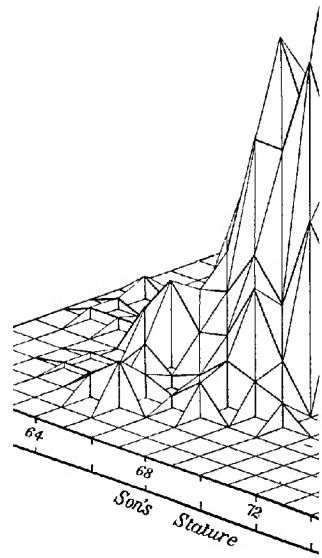
11.3.—Frequency-surface for Stature of
of th
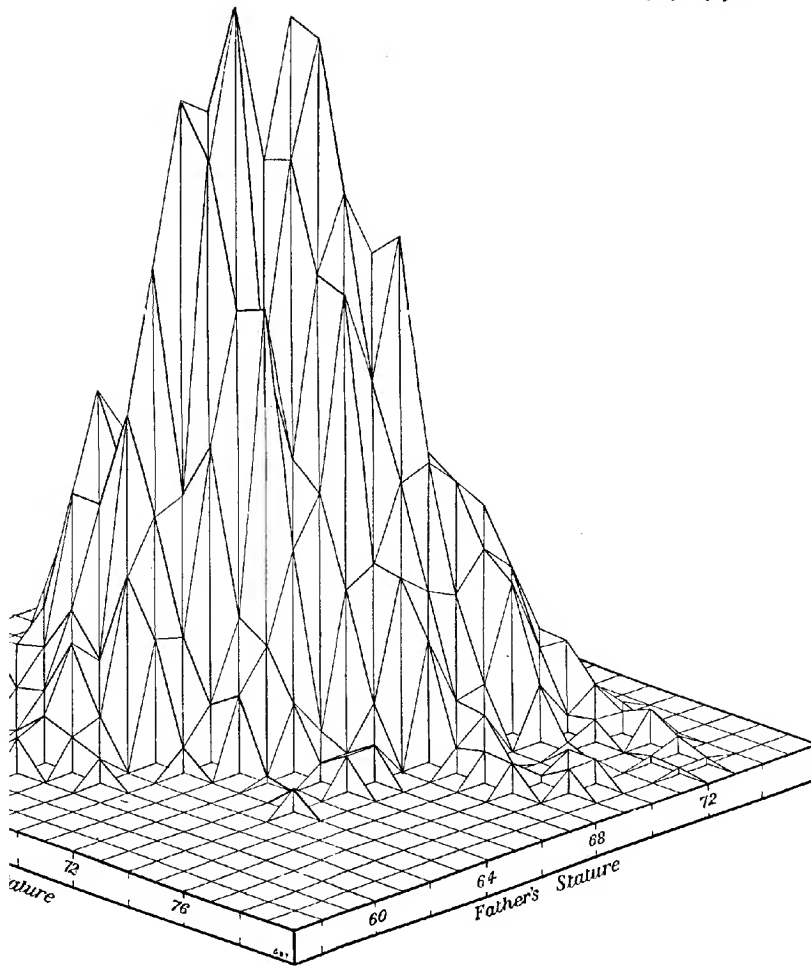
e for Stature of Father and Stature of Son (data of Table 11.3). The surface is approximately of the ideal symmetrical or "normal" form.

The Scatter Diagram.

11.9. There is another method of representing bivariate data graphically which is particularly useful for ungrouped data. Take, for instance, the data of Table 11.7, giving the index-numbers of prices of animal feeding-stuffs and home-grown oats for each month of the years 1931–35. There are only 60 pairs of values, and the data cannot be grouped into a frequency-distribution with class-intervals of reasonable size without
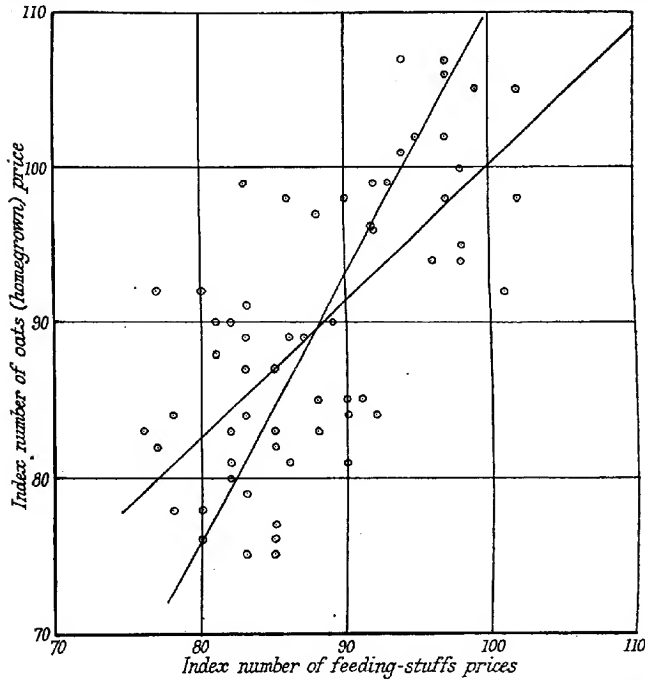
Fig. 11.4.—Scatter Diagram of Index-numbers of Prices of (1) Animal Feeding-stuffs and (2) Home-grown Oats (Table 11.7). For the meaning of the straight lines, see Example 11.1, page 217.

giving rise to irregular frequencies. We may, however, proceed as follows :—

On squared paper take two axes at right angles, one axis corresponding to the variable X and the other to the variable Y (see fig. 11.4). To each member of the universe there will correspond a pair of values X, Y, which in turn will correspond to a point whose abscissa on the diagram is X and whose ordinate is Y. · Thus the universe, when represented in this way, will give a swarm of points on the diagram, and we can interpret the ways in which these points cluster or scatter as properties of the relationship

between the two variables. Fig. 11.4 shows the data of Table 11.7 plotted in this way. It will be observed that the points tend to distribute themselves so that high and low values of X correspond to high and low values of Y respectively.

Such a figure is called a **scatter diagram.**

11.10. We can also represent a grouped bivariate frequency table on a scatter diagram, though less satisfactorily and with some labour. For this purpose axes are taken as before and abscissæ and ordinates drawn to correspond to the divisions of the frequency table. The diagram will then be divided into compartments corresponding to the compartments of the table. In each compartment we place a number of dots equal to the frequency in the corresponding compartment of the table. We have, as a rule, no guide as to the disposition of these dots within their respective cells, and hence it is usual to place them in some symmetrical arrangement so that they are, as nearly as may be, spread uniformly through the cells.

The difficulty of inserting the dots when the frequencies are large will be obvious, and, in fact, such a scatter diagram rarely tells us more than we can see from an inspection of the table itself. In contrast to this, the scatter diagram of the data of Table 11.7 gives a much better picture of the dependence of the two variates than can be obtained by mere inspection of the ungrouped data of the table.

11.11. It is clear that a correlation table may be treated by the methods discussed in Chapter 5, which are applicable to all contingency tables, however formed. But the coefficient of contingency merely tells us whether two variables are related, and if so, how closely. The methods we shall now discuss go much further than this. The numerical character of the variates and the arrangement of the correlation table in class-intervals of equal widths enable us to approach the problem of investigating the relationship between the variates with additional precision.

11.12. If the two variates in a contingency table are independent, the distributions in parallel arrays are similar **(5.18)** ; hence their averages and dispersions, *i.e.* their means and standard deviations, must be the same. In general they will not be the same, and we are thus led to inquire into the relation between the values of the means and standard deviations in different arrays and the departure of the distribution from complete independence.

11.13. The mean is the most important constant, in general, and for the present we shall concentrate our attention upon it. Although the values in arrays are scattered about their respective means, it is in most cases profitable to inquire how the means of arrays are related ; this will throw a good deal of light on the important question whether high values of one variate show any tendency to be associated, on the average, with high values of the other variate.

If possible, we also wish to know how great a divergence of one variate from its mean is associated with a given divergence of the other, and to obtain some idea of how closely the relation is usually fulfilled.

Lines of Regression.

11.14. Let us then consider the means of arrays. Let OX, OY be two axes at right angles representing the scales of the two variates. As in the case of the scatter diagram we can plot the positions of the means ; for

example, if the mean of a row whose variate value is centred at y_1 is m_1, we can plot the point whose abscissa is m_1 and whose ordinate is y_1. There will thus be one point corresponding to each row and one to each column. In practice, to distinguish the two, the means of rows are denoted by small circles and the means of columns by small crosses. Fig. 11.8 shows such a diagram drawn for the data of Table 11.3.

The means of rows and the means of columns will, in general, lie more or less closely round smooth curves. For example, in fig. 11.8 they lie, very approximately, on straight lines, RR and CC in the figure. Such curves are said to be **curves of regression,** and their equations with reference to the axes OX and OY are called **regression equations.** If the lines of regression are straight, the regression is said to be **linear.** In the contrary case it is said to be **curvilinear.**

11.15. The term " regression " is not a particularly happy one from the etymological point of view, but it is so firmly embedded in statistical literature that we make no attempt to replace it by an expression which would more suitably express its essential properties. It was introduced by Galton in connection with the inheritance of stature. Galton found that the sons of fathers who deviate x inches from the mean height of all fathers themselves deviate from the mean height of all sons by less than x inches, *i.e.* there is what Galton called a "regression to mediocrity." In general, the idea ordinarily attached to the word "regression" does not touch upon this connotation, and it should be regarded merely as a convenient term.

11.16. If two variates are independent, their regression lines are straight and at right angles, the means of rows lying on a line parallel to the axis OY and the means of columns on a line parallel to the axis OX, for the distributions in parallel arrays are similar (see fig. 11.5). In any case drawn from actual data, of course, the means might not lie exactly on straight lines, owing to fluctuations of sampling.

11.17. The cases with which the experimentalist, *e.g.* the chemist or physicist, has to deal, where the observations are all crowded closely round a single line, lie at the opposite extreme from independence. The entries fall into a few compartments only of each array, and the means of rows and of columns lie approximately on one and the same curve, like the line RR of fig. 11.6.

11.18. The ordinary cases of statistics are intermediate between these two extremes, the lines of means being neither perpendicular as in fig. 11.5, nor coincident as in fig. 11.6. One problem of the statistician is to find expressions which will suffice to describe the regression lines, either exactly or to a satisfactory degree of approximation.

In general this is a difficult problem, and the theory of curvilinear regression is as yet incomplete. We can, however, make considerable progress by confining ourselves to the cases in which the regression is linear. Cases of this kind are more frequent than might be supposed, and in other cases the means of arrays lie so irregularly, owing to the paucity of the observations, that the real nature of the regression curve is not indicated and a straight line will give as good an approximation as a more elaborate curve.

11.19. Consider the simplest case in which the means of rows lie exactly on a straight line RR (fig. 11.7). Let M_2 be the mean value of Y,

and let RR cut M_2x, the horizontal through M_2, in M. Then it may be shown that the vertical through M must cut OX in M_1, the mean of X. For, let the slope of RR to the vertical, *i.e.* the tangent of the angle M_1MR
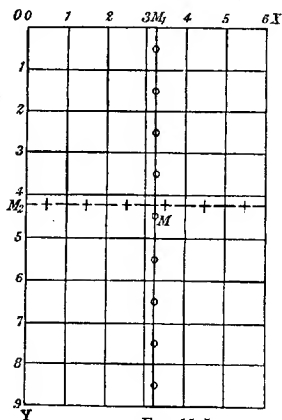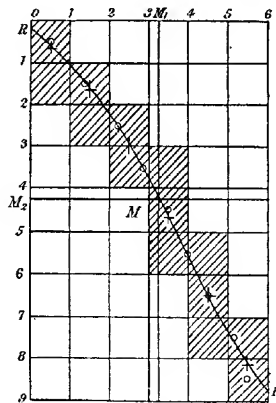
FIG. 11.5.   FIG. 11.6.
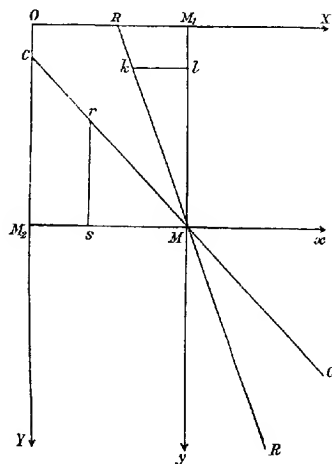
FIG. 11.7.

or ratio of kl to lM, be b_1, and let deviations from My, Mx be denoted by x and y.

Then for any one row of type y in which the number of observations is n, $S(x) = nb_1y$, and therefore for the whole table, since $S(ny) = 0$,

$S(x) = b_1 S(ny) = 0.$ M_1 must therefore be the mean of X, and M may accordingly be termed the mean of the whole distribution.

Knowing that RR passes through the mean of the distribution, we can determine it completely if we know the value of b_1.

For any one row we have

$$S(xy) = yS(x) = nb_1 y^2$$

Therefore for the whole table

$$S(xy) = b_1 S(y^2) n = N b_1 \sigma_y^2$$

Let us write

$$p = \frac{1}{N} S(xy) \qquad . \qquad . \qquad . \qquad . \quad \textbf{(11.1)}$$

Then

$$b_1 = \frac{p}{\sigma_y^2} \quad . \qquad . \qquad . \qquad . \qquad . \quad \textbf{(11.2)}$$

Similarly, if CC be the line on which lie the means of columns and b_2 is the slope to the horizontal,

$$b_2 = \frac{p}{\sigma_x^2} \quad . \qquad . \qquad . \qquad . \qquad . \quad \textbf{(11.3)}$$

Now let us define

$$r = \frac{p}{\sigma_x \sigma_y} = \frac{S(xy)}{\sqrt{S(x^2)S(y^2)}} \qquad . \qquad . \qquad . \quad \textbf{(11.4)}$$

Then

$$b_1 = r\frac{\sigma_x}{\sigma_y} \quad \text{and} \quad b_2 = r\frac{\sigma_y}{\sigma_x} \qquad . \qquad . \qquad . \quad \textbf{(11.5)}$$

and the equations of RR and CC, referred to the centre of the distribution, are

$$x = r\frac{\sigma_x}{\sigma_y} y \quad \text{and} \quad y = r\frac{\sigma_y}{\sigma_x} x \qquad . \qquad . \qquad . \quad \textbf{(11.6)}$$

and, referred to the origin 0,

$$(X - M_1) = \frac{r\sigma_x}{\sigma_y}(Y - M_2), \qquad Y - M_2 = \frac{r\sigma_y}{\sigma_x}(X - M_1) \quad . \quad \textbf{(11.7)}$$

11.20. Let us now proceed to the case when the means of arrays are not situated on a straight line. This we shall treat by finding the next best thing—straight lines which are the closest fit to the means.

The expression " closest fit," as applied to the fitting of curves to points, is one which we deal with at length in Chapter 17, and it is only necessary to say at this stage that the straight line RR of closest fit to the means of rows, i.e.

$$x = a_1 + b_1 y$$

will be determined by evaluating a_1 and b_1 so as to make the expression

$$E = S\{x - (a_1 + b_1 y)\}^2$$

(that is, the sum of the squares of the horizontal distances of the points representing the observations from RR) a minimum. Here x and y, as before, denote deviations from the respective means of X and Y, and the summation is taken over all values of x and y.

We have, expanding E,

$$E = S(a_1{}^2) - 2S\{a_1(x - b_1 y)\} + S(x - b_1 y)^2$$

The second term on the right vanishes, since $S(x) = S(y) = 0$, and hence

$$E = S(a_1{}^2) + S(x - b_1 y)^2$$

Now a_1 and b_1 can be chosen independently, and hence E is a minimum only if $S(a_1{}^2) = 0$, *i.e.*

$$a_1 = 0 \quad . \quad . \quad . \quad . \quad . \quad (11.8)$$

Thus the line of closest fit goes through the mean of the distribution. Hence,

$$
\begin{aligned}
E &= S(x - b_1 y)^2 \\
&= S(x^2) - 2b_1 S(xy) + b_1{}^2 S(y^2) \\
&= S(y^2)\left\{ b_1{}^2 - 2b_1 \frac{S(xy)}{S(y^2)} + \frac{S(x^2)}{S(y^2)} \right\} \\
&= S(y^2)\left[\left\{ b_1 - \frac{S(xy)}{S(y^2)} \right\}^2 + \frac{S(x^2)}{S(y^2)} - \left\{ \frac{S(xy)}{S(y^2)} \right\}^2 \right] \quad . \quad . \quad (11.9)
\end{aligned}
$$

This is a minimum when the first term (a square) is zero, *i.e.* when

$$b_1 = \frac{S(xy)}{S(y^2)} \quad . \quad . \quad . \quad . \quad (11.10)$$

which is the same as equation (11.2).

We may show similarly that the line of closest fit CC, given by

$$y = a_2 + b_2 x$$

has

$$a_2 = 0, \qquad b_2 = \frac{S(xy)}{S(x^2)},$$

which is the same as equation (11.3).

If we regard the equation

$$x = a_1 + b_1 y$$

as one for estimating x from y, we may take $x - a_1 - b_1 y$ as an error of estimation, and E will then be the sum of the squares of such errors. The condition that E is a minimum is then equivalent to the condition that the sum of squares of errors of estimation shall be a minimum. This is one form of the so-called " Principle of Least Squares " (see Chapter 17).

11.21. Equations (11.6) and (11.7) are thus of general application. If the regression is exactly linear they give the lines of regression. If the regression departs from linearity, either owing to sampling effects or owing to real divergences, they give the " best " straight regression lines which the data admit. We may regard the equations as either (a) equations for estimating an individual x from its associated y (or y from its associated x) in such a way that the sum of squares of errors of estimation is a minimum ;

or (*b*) equations for estimating the *mean* of the *x*'s associated with a particular *y* (or the mean of *y*'s associated with a particular *x*) in such a way that the sum of the squares of errors of estimation is a minimum, each mean being counted proportionately to the number of observations on which it is based.

Coefficient of Correlation.

11.22. The coefficient *r* defined in equation (11.4) is of very great importance. It is called the **coefficient of correlation.**

r cannot exceed + 1 or be less than − 1.

For, from equation (11.9) we see that the value of *E* is

$$S(x - b_1 y)^2 = S(x^2) - \frac{\{S(xy)\}^2}{S(y^2)} = S(x^2)\{1 - r^2\} \qquad . \quad (11.11)$$

But *E* is the sum of a number of squares and cannot be negative. Hence,

$$1 - r^2 \geqslant 0$$

which proves the result.

If *r* = + 1, the regression equations are identical, as may be seen from equations (11.6), and hence the lines *RR* and *CC* coincide. In this case it follows from (11.11) that for all pairs of values of the variates

$$x - b_1 y = 0$$

i.e. all values lie on a single straight line. Thus to one value of *x* there
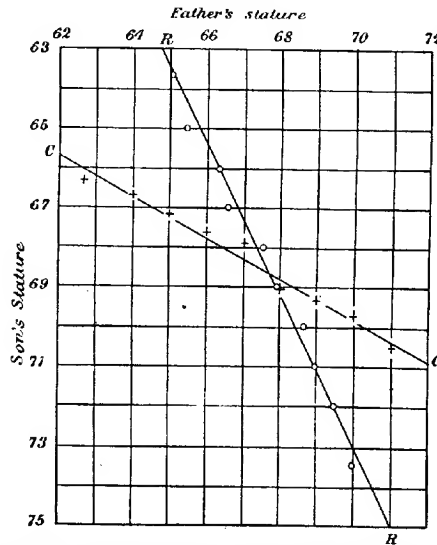
FIG. 11.8.—Correlation between Stature of Father and Stature of Son (Table 11.3): means of rows shown by circles and means of columns by crosses: *r* = + 0·51.

corresponds one, and only one, value of y.   This is the case we mentioned in **11.17**, and since high values of x correspond to high values of y, the variables may be said to be perfectly positively correlated.

Similarly, if $r = -1$, the pairs of values all lie on a single straight line as before, but high values of one will be associated with low values of the
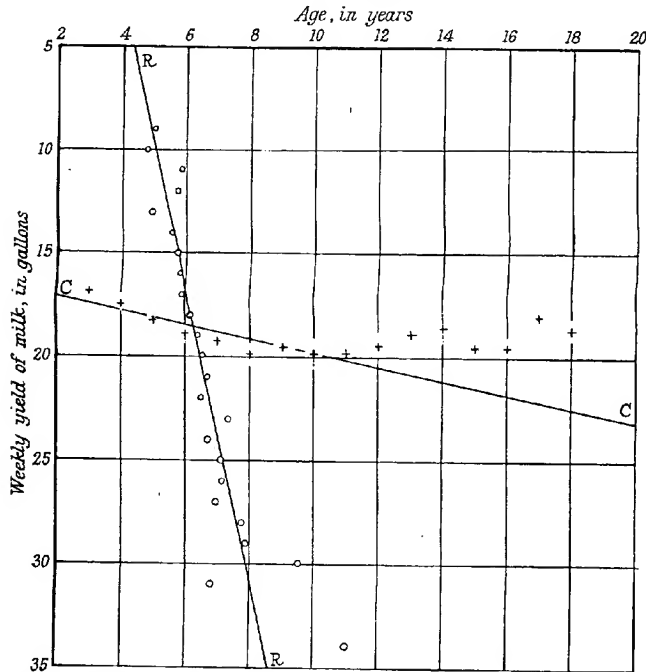
Fig. 11.9.—Correlation between Age and Weekly Yield of Milk from Cows (Table 11.4): means of rows shown by circles and means of columns by crosses: $r = +0.22$.

other.   In this case we can say that the variates are perfectly negatively correlated.

Finally, if the variates are independent, r is zero, for b_1 and b_2 are zero, and the lines of regression are parallel to OX and OY.   It does *not* follow, however, that if r is zero the variates are independent; the fact that r is zero implies only that the means of arrays lie *scattered around* two straight lines which do not exhibit any definite trend away from the horizontal or the vertical as the case may be.   Two variates for which r is zero may, however, be spoken of as *uncorrelated*.   Table 11.6 will serve as a case where the variates are almost uncorrelated but by no means independent, r being very small (-0.014) (see fig. 11.10), but the coefficient of contingency C (for the grouping of Exercise 11.3) 0.47.   Figs. 11.8 and

11.9 are drawn from the data of Tables 11.3 and 11.4, for which r has the values $+0.51$ and $+0.22$ respectively. The student should study such tables and diagrams closely, and endeavour to accustom himself to estimating the value of r from the general appearance of the table.
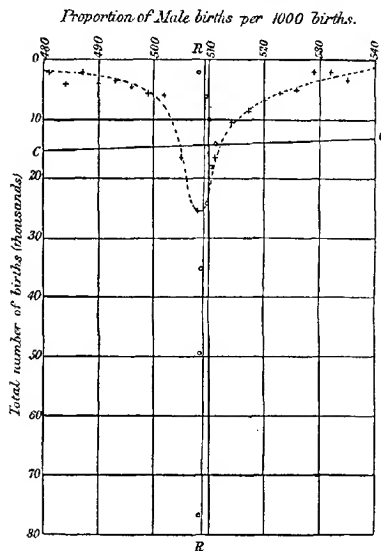
FIG. 11.10.—Correlation between Births in a Registration District and Proportion of Male Births per Thousand of All Births in England and Wales, 1881-90 (Table 11.6): means of rows shown by circles and means of columns by crosses: $r = -0.014$.

Coefficients of Regression.

11.23. The two quantities

$$b_1 = \frac{r\sigma_x}{\sigma_y}, \qquad b_2 = \frac{r\sigma_y}{\sigma_x}$$

are called **coefficients of regression,** b_1 being the regression of x on y, or deviation in x corresponding on the average to a unit change in y, and b_2 being similarly the regression of y on x.

The coefficient of correlation is always a pure number, but the coefficients of regression are only pure numbers if the variates are the same in kind; for they depend on the ratio $\dfrac{\sigma_x}{\sigma_y}$, and consequently on the units in which x and y are measured.

. Since r is not greater than unity, one of the coefficients of regression is less than unity; but the other may be greater than unity, if $\dfrac{\sigma_x}{\sigma_y}$ or $\dfrac{\sigma_y}{\sigma_x}$ be large.

11.24. The two standard deviations,

$$s_x = \sigma_x \sqrt{1 - r^2}, \qquad s_y = \sigma_y \sqrt{1 - r^2}$$

are of considerable importance. It follows from (11.11) that s_x is the standard deviation of $(x - b_1 y)$, and similarly s_y is the standard deviation of $(y - b_2 x)$. Hence we may regard s_x and s_y as the standard errors (root-mean-square errors) made in estimating x from y and y from x by the respective regression equations

$$x = b_1 y, \qquad y = b_2 x$$

s_x may also be regarded as a kind of average standard deviation of a row about RR, and s_y as an average standard deviation of a column about CC. In an ideal case, where the regression is truly linear and the standard deviations of all parallel arrays are equal, a case to which the distribution of Table 11.3 is a rough approximation,[1] s_x is the standard deviation of the x-array and s_y the standard deviation of the y-array. Hence s_x and s_y are sometimes termed the " standard deviations of arrays."

Calculation of the Coefficient of Correlation.

11.25. We now proceed to the arithmetical work involved in calculating the correlation coefficient.

For this purpose we use the formula (11.4), *i.e.*

$$r = \frac{S(xy)}{N \sigma_x \sigma_y} = \frac{S(xy)}{\sqrt{S(x^2) S(y^2)}}$$

The calculation of $S(x^2)$, or σ_x, and of $S(y^2)$, or σ_y, proceeds exactly as in Chapter 8. The only expression of a novel type is the quantity $\frac{1}{N} S(xy)$, which we may call the first product-moment of the distribution.[2] As in the case of univariate distributions, the form of the arithmetic is slightly different according as the observations are grouped or ungrouped.

 11.26. Our work is greatly simplified by the use of devices similar to those employed in calculating the means and other moments of univariate distributions.

(*a*) We take working means for the two variates, obtained by inspection, and transfer our moments to those about the means after the bulk of the arithmetic has been performed. For the first product-moment

---

[1] Arrays in which the standard deviations are equal are sometimes said to be "homoscedastic"; in the contrary case "heteroscedastic."

[2] In generalisation of the definition of moments of a univariate distribution in Chapter 9 we may define the product-moments of a bivariate universe as

$$\mu_{rs} = \frac{1}{N} S(f x^r y^s)$$

where f is the frequency. This gives us

$$\mu_{11} = \frac{1}{N} S(fxy)$$

the quantity we have called p in equation (11.1).

we have, in fact, if ξ, η are the deviations from the working means and $\bar{\xi}$, $\bar{\eta}$ the deviations of the true means from the working means :

$$\xi = x + \bar{\xi}, \qquad \eta = y + \bar{\eta}$$

Hence,

$$\xi\eta = xy + \bar{\xi}y + x\bar{\eta} + \bar{\xi}\bar{\eta}$$

Summing for all members of the universe, since $S(\bar{\xi}y) = \bar{\xi}S(y) = 0$ and similarly $S(x\bar{\eta}) = 0$, x and y being deviations from the true means,

$$S(\xi\eta) = S(xy) + N\bar{\xi}\bar{\eta}$$

Hence,

$$S(xy) = S(\xi\eta) - N\bar{\xi}\bar{\eta}. \qquad . \qquad . \qquad . \quad (11.12)$$

This gives us the product-moment about the true means in terms of the product-moment about the working means and the deviations of the true means from the working means.

(b) As a check on the rather heavy arithmetic which is frequently involved, it is advisable to use a method similar to that of **8.10**. We have

$$S(\xi+1)(\eta+1) = S(\xi\eta) + S(\xi) + S(\eta) + N \quad . \qquad . \quad (11.13)$$

If, therefore, we calculate $S(\xi+1)(\eta+1)$ as well as $S(\xi\eta)$, we shall have in the above equation a check on the accuracy of our work.

(c) We take the class-intervals as units and transfer to other units afterwards as desired.

Example 11.1, Table 11.8.—Let us investigate the correlation and regressions of the variates of Table 11.7, the data of which are ungrouped. The variates are (1) the price index-number of animal feeding-stuffs, X, and (2) the price index-number of home-grown oats, Y. The values of the variates themselves are shown in columns 2 and 3 of Table 11.8. We take a working mean at $X = 90$ and $Y = 90$, and the deviations from these values are shown in columns 4 and 5. The remaining columns 6 to 13 give the squares and product of the deviations together with the various auxiliary quantities used for checking purposes. Finally, the various sums are shown at the bottom of the table.

In practice it is as well to show the negative values which may occur in columns 4, 5, 6, 7, 12 and 13 (particularly the last two) in a separate column, so as to facilitate addition and avoid mistakes. We have refrained from this course for convenience of printing.

As check on the arithmetic we have :

$$-118 = S(\xi) = S(\xi+1) - N = -58 - 60$$

$$2924 = S(\xi+1)^2 = S(\xi^2) + 2S(\xi) + N = 3100 - 236 + 60$$

etc., and

$$2493 = S(\xi+1)(\eta+1) = S(\xi\eta) + S(\xi) + S(\eta) + N$$
$$= 2565 - 118 - 14 + 60$$
$$= 2493$$

TABLE 11.8.—*Correlation between Monthly Index-numbers of Prices of* (1) *Animal Feeding-stuffs and* (2) *Home-grown Oats in Years 1931-35.*

1. Month.	2. X.	3. Y.	4. ξ.	5. η.	6. $\xi+1$.	7. $\eta+1$.	8. ξ^2.	9. $(\xi+1)^2$.	10. η^2.	11. $(\eta+1)^2$.	12. $\xi\eta$.	13. $(\xi+1)(\eta+1)$.
1931 Jan.	78	84	−12	− 6	−11	− 5	144	121	36	25	72	55
Feb.	77	82	−13	− 8	−12	− 7	169	144	64	49	104	84
Mar.	85	82	− 5	− 8	− 4	− 7	25	16	64	49	40	·28
Apr.	88	85	− 2	− 5	− 1	− 4	4	1	25	16	10	4
May	87	89	·· 3	− 1	− 2	—	9	4	1	—	3	—
June	82	90	− 8	—·	—	1	64	49	—	1	—	− 7
July	81	88	− 9	− 2	− 8	− 1	81	64	4	1	18	8
Aug.	77	92	−13	2	−12	3	169	144	4	9	−26	−36
Sept.	76	83	−14	− 7	−13	− 6	196	169	· 49	36	98	78
Oct.	83	89	− 7	− 1	− 6	—	49	36	1	—	7	—
Nov.	97	98	7	8	8	9	49	64	64	81	56	72
Dec.	93	99	3	9	4	10	9	16	81	100	27	40
1932 Jan.	95	102	5	12	6	13	25	36	144	169	60	78
Feb.	97	102	7	12	8	13	49	64	144	169	84	104
Mar.	102	105	12	15	13	16	144	169	225	·256	180	208
Apr.	99	105	9	15	10	16	81	100	225	256	135	160
May	97	107	7	17	8	18	49	64	289	324	119	144
June	94	107	4	17	5	18	16	25	289	324	68	90
July	94	101	4	11	5	12	16	25	121	144	44	60
Aug.	97	106	7	16	8	17	49	64	256	289	112	136
Sept.	92	96	2	6	3	7	4	9	36	49	12	21
Oct.	89	90	− 1	—	·	1	1	—	—	1	—	—
Nov.	90	85	—	− 5	1	·· 4	—	1	25	16	—	− 4
Dec.	90	81	—·	− 9	1	− 8	—	1	81	64	—	− 8
1933 Jan.	92	84	2	− 6	3	− 5	4	9	36	25	−12	−15
Feb.	91	85	1	− 5	2	− 4	1	4	25	16	− 5	− 8
Mar.	90	84	··	·· 6	1	− 5	—	1	36	25	—	− 5
Apr.	86	81	− 4	− 9	− 3	− 8	16	9	81	64	36	24
May	85	76	− 5	−14	− 4	−13	25	16	196	169	70	52
June	85	77	− 5	−13	− 4	−12	25	16	169	144	65	48
July	85	75	− 5	−15	− 4	−14	25	16	225	196	75	56
Aug.	83	79	− 7	−11	− 6	−10	49	36	121	100	77	60
Sept.	80	78	−10	−12	− 9	−11	100	81	144	121	120	99
Oct.	78	78	−12	−12	−11	−11	144	121	144	121	144	121
Nov.	80	76	−10	−14	− 9	−13	100	81	196	169	140	117
Dec.	83	75	− 7	−15	− 6	−14	49	36	225	196	105	84
1934 Jan.	82	80	− 8	−10	·· 7	·· 9	64	49	100	81	80	63
Feb.	83	91	− 7	1	− 6	2	49	36	1	4	− 7	−12
Mar.	85	87	− 5	− 3	− 4	− 2	25	16	9	4	15	8
Apr.	83	84	− 7	− 6	− 6	− 5	49	36	36	25	42	30
May	82	81	− 8	− 9	− 7	− 8	64	49	81	64	72	56
June	85	83	− 5	− 7	·· 4	·· 6	25	16	49	36	35	24
July	88	83	− 2	− 7	− 1	·· 6	4	1	49	36	14	6
Aug.	101	92	11	2	12	3	121	144	4	9	22	36
Sept.	102	98	12	8	13	9	144	169	64	81	96	117
Oct.	98	94	8	4	9	5	64	81	16	25	32	45
Nov.	96	94	6	4	7	5	36	49	16	25	24	35
Dec.	98	95	8	5	9	6	64	81	25	36	40	54
1935 Jan.	98	100	8	10	9	11	64	81	100	121	80	99
Feb.	92	99	2	9	3	10	4	9	81	100	18	30
Mar.	92	96	2	6	3	7	4	9	36	49	12	21
Apr.	90	98	—	8	1	9	—	1	64	81	—	9
May	88	97	− 2	7	− 1	8	4	1	49	64	−14	− 8
June	86	98	− 4	8	− 3	9	16	9	64	81	−32	−27
July	83	99	− 7	9	− 6	10	49	36	81	100	−63	−60
Aug.	80	92	−10	2	− 9	3	100	81	4	9	−20	−27
Sept.	81	90	− 9	—	− 8	1	81	64	··	1	··	− 8
Oct.	86	89	− 4	− 1	− 3	—	16	9	1	—	4	—
Nov.	83	87	− 7	− 3	·· 6	− 2	49	36	9	4	21	12
Dec.	82	83	− 8	− 7	− 7	− 6	64	49	49	36	56	42
Total	—	—	−118	−14	−58	46	3100	2924	4814	4846	2565	2493

We have, then, about the working means :

$$\bar{\xi} = -\frac{118}{60} = -1\cdot9667$$

$$\bar{\eta} = -\frac{14}{60} = -0\cdot2333$$

$$\sigma_x^2 = \frac{3100}{60} - \bar{\xi}^2 = 47\cdot7989, \qquad \sigma_x = 6\cdot914$$

$$\sigma_y^2 = \frac{4814}{60} - \bar{\eta}^2 = 80\cdot1789, \qquad \sigma_y = 8\cdot954$$

$$p = \frac{S(xy)}{N} = \frac{S(\xi\eta)}{N} - \bar{\xi}\bar{\eta} = 42\cdot75 - 0\cdot4589 = 42\cdot2911$$

$$r = \frac{p}{\sigma_x\sigma_y} = \frac{42\cdot2911}{61\cdot9080} = +0\cdot68$$

Further, working the regressions in the way best to avoid errors in rounding off,

$$b_1 = \frac{p}{\sigma_y^2} = 0\cdot527$$

$$b_2 = \frac{p}{\sigma_x^2} = 0\cdot885$$

Thus the correlation coefficient is 0·68, and the regression equations, referred to the means, are :

$$x = 0\cdot527y$$
$$y = 0\cdot885x$$

If we prefer to express these equations with origin at $X = 0$, $Y = 0$, we have :

$$X - (90 - 1\cdot97) = X - 88\cdot03 = 0\cdot527(Y - 89\cdot77)$$
$$Y - (90 - 0\cdot23) = Y - 89\cdot77 = 0\cdot885(X - 88\cdot03)$$

which reduce to

$$X = 0\cdot527Y + 40\cdot72 \qquad . \qquad . \qquad . \qquad . \qquad (a)$$
$$Y = 0\cdot885X + 11\cdot86 \qquad . \qquad . \qquad . \qquad . \qquad (b)$$

The lines of regression are drawn on the scatter diagram of fig. 11.4.

The standard errors made in using these equations to estimate the index-number of oats from animal feeding-stuffs, and *vice versa*, are :

$$\sigma_x\sqrt{1 - r^2} = 5\cdot07$$
$$\sigma_y\sqrt{1 - r^2} = 6\cdot57$$

Equation (a) tells us that a rise of one point in the price index-number of oats is accompanied *on the average* by a rise of 0·527 point in the price index-number of feeding-stuffs. Similarly, equation (b) tells us that a rise of one point in the index for feeding-stuffs is accompanied *on the average* by a rise of 0·885 point in the price of oats.

It is important to note that the regression equations do not tell us

whether a variation in one variate is *caused* by a variation in the other; all we know is that the two vary together, and so far as the regression equations show, either the feeding-stuffs price may exert an influence on the oats price, or *vice versa*, or their common variation may be due to some other cause affecting both. This is only one instance of a difficulty which pervades the theory of correlation and regression, namely, that of *interpreting* results in terms of causal factors.

Example 11.2, Table 11.9.—We now consider an example based on grouped data. In this we have omitted the auxiliary quantities necessary for checking in order to save space.

(Unpublished data ; measurements by G. U. Yule.) The two variables are (1) X, the length of a mother-frond of duckweed (*Lemna minor*); (2) Y, the length of the daughter-frond. The mother-frond was measured when the daughter-frond separated from it, and the daughter-frond when its first daughter-frond separated. Measures were taken from camera drawings made with the Zeiss-Abbé camera under a low power, the actual magnification being 24 : 1. The units of length in the tabulated measurements are millimetres on the drawings.

The arbitrary origin for both X and Y was taken at 105 mm. The following are the values found for the constants of the single distributions :—

$\xi = -1 \cdot 058$ intervals $= - 6 \cdot 3$ mm. $M_1 = 98 \cdot 7$ mm. on drawing
                                                                   $= 4 \cdot 11$ mm. actual

$\sigma_x = 2 \cdot 828$ intervals $= 17 \cdot 0$ mm. on drawing $= 0 \cdot 707$ mm. actual

$\bar{\eta} = -0 \cdot 203$ interval $= - 1 \cdot 2$ mm. $M_2 = 103 \cdot 8$ mm. on drawing
                                                                   $= 4 \cdot 32$ mm. actual

$\sigma_y = 3 \cdot 084$ intervals $= 18 \cdot 5$ mm. on drawing $= 0 \cdot 771$ mm. actual

To calculate $S(\xi\eta)$ the value of $\xi\eta$ is first written in every compartment of the table against the corresponding frequency, treating the class-interval as unit. In Table 11.9 frequencies are shown in ordinary type and the values of $\xi\eta$ in heavy type. In making these entries the sign of the product may be neglected, but it must be remembered that this sign will be positive in the upper left-hand and lower right-hand quadrants, and negative in the two others. The frequencies are then collected, according to the magnitude and sign of $\xi\eta$, in columns 2 and 3 of Table 11.10. When columns 2 and 3 are completed they should be checked to see that no frequency has been dropped, which may readily be done by adding together the totals of the two columns and the frequency in the 8th row and 8th column of Table 11.9 (the row and column for which $\xi\eta = 0$), care being taken not to count twice the frequency in the compartment common to the two. This grand total must clearly be equal to N, the total number of observations, which in this case is 266. The numbers in column 4 are given by deducting the entries in column 3 from those in column 2. The totals so obtained are multiplied by $\xi\eta$ (column 1) and the products entered in column 5 or 6 according to sign. The algebraic sum of these totals gives

$$S(\xi\eta) = +1519 \cdot 5$$

TABLE 11.9.—THEORY OF CORRELATION: *Example 11.2.—Correlation between* (1) *Length of Mother-frond*, (2) *Length of Daughter-frond, in Lemna minor.* [Unpublished data; G. U. Yule.] (The frequencies are the figures printed in ordinary type. The numbers in heavy type are the deviation-products $(\xi\eta)$.)

In each cell, the upper number (ordinary type) is the frequency; the lower number (**heavy type**) is the deviation-product.

(2) Length of daughter-frond	60-66	66-72	72-78	78-84	84-90	90-96	96-102	102-108	108-114	114-120	120-126	126-132	132-138	138-144	144-150	150-156	156-162	162-168	Total
60-66	—	2 / **42**	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2
66-72	—	—	—	1 / **24**	2 / **18**	—	—	—	—	—	—	—	—	—	—	—	—	—	2
72-78	—	—	1 / **25**	3 / **20**	1 / **15**	—	2 / **5**	2 / **0**	0·5 / **5**	—	—	—	—	—	—	—	—	—	9·5
78-84	—	4 / **24**	2 / **20**	5 / **16**	6 / **12**	5 / **8**	—	3 / **0**	1·5 / **4**	—	—	—	—	—	—	—	—	—	26·5
84-90	2 / **21**	—	3 / **15**	4·5 / **12**	2 / **9**	2 / **6**	3 / **3**	5 / **0**	2 / **3**	2 / **6**	2 / **9**	—	—	—	—	—	—	—	27·5
90-96	1 / **14**	1 / **12**	3·5 / **10**	5 / **8**	6·5 / **6**	4·5 / **4**	4 / **2**	4 / **0**	3 / **2**	—	1 / **6**	—	—	—	—	—	—	—	33·5
96-102	—	—	—	8 / **4**	4·5 / **3**	7 / **2**	—	1 / **0**	3 / **1**	1 / **2**	—	—	—	—	—	—	—	—	24·5
102-108	—	—	1 / **0**	8 / **0**	7·5 / **0**	7 / **0**	3·5 / **0**	9 / **0**	3 / **0**	—	—	—	—	—	—	—	—	—	39
108-114	—	—	—	—	4 / **3**	5 / **2**	5·5 / **1**	6 / **0**	—	2 / **2**	—	—	—	—	—	—	—	—	22·5
114-120	—	—	—	—	1 / **6**	5 / **4**	4·5 / **2**	5·5 / **0**	4 / **2**	—	1 / **6**	—	1 / **10**	—	—	—	—	—	22
120-126	—	—	—	—	2 / **9**	1 / **6**	3 / **3**	2 / **0**	3 / **3**	3 / **6**	7 / **9**	2 / **12**	—	—	—	—	—	—	23
126-132	—	—	—	—	—	1 / **8**	—	3 / **0**	1 / **4**	3 / **8**	2 / **12**	2 / **16**	3 / **20**	—	—	—	—	—	15
132-138	—	—	—	—	—	1 / **10**	—	1 / **0**	—	2 / **10**	2 / **15**	—	—	2 / **30**	—	—	—	—	8
138-144	—	—	—	—	—	—	—	—	1 / **6**	2 / **12**	—	—	1 / **30**	1 / **36**	—	—	—	1 / **60**	6
144-150	—	—	—	—	—	—	—	—	—	—	—	1 / **28**	—	—	—	—	—	—	1
150-156	—	—	—	—	—	—	—	—	—	—	1 / **24**	—	—	—	—	—	—	—	1
156-162	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1 / **63**	—	—	—	1
162-168	—	—	—	—	—	—	—	—	—	—	—	1 / **40**	—	—	—	—	—	—	1
Total	3	7	10·5	34·5	36·5	38·5	25·5	41·5	22	15	16	6	5	3	1	—	—	1	206

(1) Length of mother-frond (mm. of camera drawing enlarged 24 : 1).

TABLE 11.10.

1.	2.	3.	4.	5.	6.
	Frequencies.			Products.	
$\xi\eta$.			Total.		
	+ Quadrants.	− Quadrants.		+	−
1	—	8·5	− 8·5	—	8·5
2	17	13·5	+ 3·5	7	—
3	10·5	9	+ 1·5	4·5	—
4	13·5	6·5	+ 7	28	—
5	2	0·5	+ 1·5	7·5	—
6	13·5	5	+ 8·5	51	—
8	13	1	+12	96	—
9	9	4	+ 5	45	—
10	6·5	1	+ 5·5	55	—
12	17·5	—	+17·5	210	—
14	1	—	+ 1	14	—
15	6	—	+ 6	90	—
16	7	—	+ 7	112	—
18	2	—	+ 2	36	—
20	8	—	+ 8	160	—
21	2	—	+ 2	42	—
24	6	—	+ 6	144	—
25	1	—	+ 1	25	—
28	1	—	+ 1	28	—
30	3	—	+ 3	90	—
36	1	—	+ 1	36	—
40	1	—	+ 1	40	—
42	2	—	+ 2	84	—
60	1	—	+ 1	60	—
63	1	—	+ 1	63	—
Totals	145·5 49 71·5 266	49	—	+1528 − 8·5 1519·5	− 8·5

Hence, dividing by 266,

$$\frac{1}{N}S(\xi\eta) = 5\cdot712$$

$$p = 5\cdot712 - \xi\bar{\eta} = 5\cdot712 - 0\cdot215$$
$$= 5\cdot497$$

Hence,

$$r = \frac{p}{\sigma_x\sigma_y} = \frac{5\cdot497}{2\cdot828 \times 3\cdot084} = +0\cdot63$$

The regression of daughter-frond on mother-frond is 0·69 (a value which will not be affected by altering the units of measurement for both mother- and daughter-fronds, as such an alteration will affect both standard deviations equally). Hence, the regression equation giving the

average actual length (in millimetres) of daughter-fronds for mother-fronds
of actual length X is
$$Y = 1 \cdot 48 + 0 \cdot 69X$$

We leave it to the student to work out the second regression equation
giving the average length of mother-fronds for daughter-fronds of length Y,
and to check the whole work by a diagram showing the lines of regression
and the means of arrays for the central portion of the table.

Example 11.3, Table 11.2.—The following device is frequently useful,
and saves a considerable amount of labour in calculating the product
term $S(xy)$.

We have:
$$S(x - y)^2 = S(x^2) - 2S(xy) + S(y^2) \qquad . \qquad . \qquad . \quad \text{(i)}$$
and
$$S(x + y)^2 = S(x^2) + 2S(xy) + S(y^2) \qquad . \qquad . \qquad . \quad \text{(ii)}$$

Hence, knowing $S(x^2)$ and $S(y^2)$, we can find $S(xy)$ if we know either
$S(x - y)^2$ or $S(x + y)^2$. These quantities are often easier to calculate than
$S(xy)$ itself.

Consider the data of Table 11.2. In the usual way, taking a working
mean centred in the intervals $X = 25-$ years, $Y = 25-$ years, we have, in
units of five years:

$$\bar{\xi} = + 0 \cdot 2924 \qquad \bar{\eta} = - 0 \cdot 2353$$
$$S(\xi^2) = 9708 \qquad S(\eta^2) = 7090$$
$$\sigma_x = 1 \cdot 730 \qquad \sigma_y = 1 \cdot 481$$

Now the value of $\xi - \eta$ is constant down diagonals which run from the
top left hand to the bottom right hand of the table. In fact, for the
principal diagonal, running from $X = 15-$, $Y = 15-$ through $X = 20-$, $Y = 20-$,
etc., $\xi - \eta = 0$. For the diagonal above this, running from $X = 20-$, $Y = 15-$
through $X = 25-$, $Y = 20-$, etc., $\xi - \eta = 1$, and so on.

Let us then find the diagonal totals. We find :

$\xi - \eta$.	Frequency in diagonal.
-3	4
-2	34
-1	280
0	1398
1	1051
2	263
3	73
4	31
5	12
6	5
7	2
	3153

The total is the total frequency, which gives a check on the work.

The value of $S(\xi - \eta)^2$ for the whole table is then obtained from the
above table by squaring the values in the left-hand column, multiplying

by the corresponding frequency in the right-hand column and adding. We get

$$S(\xi - \eta)^2 = (9 \times 4) + (4 \times 34) + (1 \times 280) + \ldots + (49 \times 2)$$
$$= 4286$$

Hence, from (i),

$$4286 = 9708 + 7090 - 2S(\xi\eta)$$
$$\therefore \quad S(\xi\eta) = 6256$$
$$p = \frac{6256}{3153} - \xi\bar{\eta} = +2 \cdot 0529$$

whence

$$r = \frac{p}{\sigma_x \sigma_y} = +\frac{2 \cdot 0529}{1 \cdot 730 \times 1 \cdot 481} = 0 \cdot 80$$

The regression equations may now be obtained in the usual manner.

In the above work we chose equation (i) in preference to equation (ii) because the frequencies are seen by inspection to run mainly from the top left hand to the bottom right hand of the table. Had they run from the top right hand to the bottom left hand we should probably have found it better to use equation (ii).

11.27. The student should be careful to remember the following points in working :—

(1) To give $S(\xi\eta)$ and $\xi\bar{\eta}$ their correct signs in finding the true mean deviation product p.

(2) To express σ_x and σ_y in terms of the class-interval as a unit, in the value of $r = p/\sigma_x\sigma_y$, for these are the units in terms of which p has been calculated.

(3) To use the proper units for the standard deviations (not class-intervals in general) in calculating the coefficients of regression : in forming the regression equation in terms of the absolute values of the variables, for example, as above, the work will be wrong unless means and standard deviations are expressed in the same units.

Fluctuations of Sampling.

11.28. Further, it must always be remembered that correlation coefficients, like other statistical measures, are subject to fluctuations of sampling. We shall consider this point at some length in later chapters (21 and 23), since the correlation coefficient has certain individual features which make it of special interest from the sampling point of view. We may, however, at this stage stress that if the number of observations is small, no significance can be attached to small, or even moderately large, values of r as indicating a real correlation in the universe from which the observations are drawn. For example, if $N = 36$, a value of $r = \pm 0 \cdot 5$ may be a chance result, though a very infrequent one, in sampling from an uncorrelated universe. If $N = 100$, $r = \pm 0 \cdot 3$ may similarly be a mere fluctuation of sampling, though again a very infrequent one. The student should therefore be careful in interpreting his coefficients.

Corrections for Grouping.

11.29. In this connection we may mention the question whether, in calculating the correlation coefficient from grouped data, any correction

is to be made analogous to the Sheppard correction for grouping which we have considered in the case of univariate data. In the examples considered in the foregoing we have not made such corrections.

It appears that, when the distribution is reasonably symmetrical and obeys conditions similar to those enunciated in **8.11**, page 141, we may, with advantage, correct the standard deviations σ_x, σ_y, by applying to each the formula

$$\sigma^2(\text{corrected}) = \sigma^2 - \frac{h^2}{12}$$

where h is the width of the interval. The product term $S(xy)$ needs no such correction.

We pointed out in **8.11**, however, that sampling fluctuations usually obliterate any correction for grouping unless the size of the sample is large. It may, as before, be suggested that unless $N = 1000$ or more, it is hardly worth while making the correction. For example, in Tables 11.1–11.6, Tables 11.1, 11.5 and 11.6 have a frequency less than 1000 and the corrections are not to be applied—in any case they would not be applied to Tables 11.5 and 11.6, which violate the conditions as to " tapering off."

11.30. Finally, it should be borne in mind that any coefficient, *e.g.* the coefficient of correlation or the coefficient of contingency, gives only a part of the information afforded by the original data or the correlation table. The correlation table itself, or the original data if no correlation table has been compiled, should always be given, unless considerations of space or of expense absolutely preclude the adoption of such a course.

SUMMARY.

1. A universe every member of which bears one of the values of each of two variates is said to be bivariate. If the members are grouped according to class-intervals of the two variables, we have a bivariate frequency-distribution.

2. The bivariate frequency-distribution may be represented by a frequency-surface or by a stereogram. Ungrouped data (and, less conveniently, grouped data) can be represented on a scatter diagram.

3. The means of arrays of a bivariate frequency-distribution may be represented as points by reference to a pair of rectangular axes along which are measured values of the variables. The means of rows and those of columns will in general lie respectively about two smooth curves, called lines of regression. The equations of these curves are called regression equations.[1]

4. The regression equations may be regarded as expressions for estimating from a given value of one variate the average corresponding value of the other.

5. The coefficient of correlation (product-moment correlation coefficient) between two variables X and Y is given by :

---

[1] Curvilinear regression lines, like straight regression lines, may also be defined for ungrouped data by an extension of the principle of making sums of squares of errors of estimate a minimum.

$$r = \frac{S(xy)}{\sqrt{S(x^2)S(y^2)}}$$

$$= \frac{p}{\sigma_x \sigma_y}$$

where x, y are the values of the variables measured from their respective means, and $p = \dfrac{S(xy)}{N}$.

6. The correlation coefficient r cannot be less than -1 or greater than $+1$. If $r = \pm 1$ the variables are perfectly correlated, the points corresponding to pairs of values x, y all lying on a straight line. If $r = -1$ the variables are perfectly negatively correlated, low values of one corresponding to high values of the other. If $r = +1$ the variables are perfectly positively correlated, high values of one corresponding to high values of the other.

7. The linear regression equation of X on Y (referred to axes through their respective means) is

$$x = b_1 y$$

where

$$b_1 = \frac{r\sigma_x}{\sigma_y} = \frac{p}{\sigma_y^2}$$

and that of Y on X is

$$y = b_2 x$$

where

$$b_2 = \frac{r\sigma_y}{\sigma_x} = \frac{p}{\sigma_x^2}$$

b_1 and b_2 being called coefficients of regression, or simply regressions.

8. The straight lines of regression are such that the sums of squares of errors of estimate, $S(x - b_1 y)^2$ and $S(y - b_2 x)^2$, are a minimum. If the quotients of these sums by N are denoted by s_x^2, s_y^2,

$$s_x^2 = \sigma_x^2(1 - r^2)$$
$$s_y^2 = \sigma_y^2(1 - r^2)$$

EXERCISES.

11.1. Find the correlation coefficient and the equations of regression for the following values of X and Y :—

X.	Y.
1	2
2	5
3	3
4	8
5	7

[As a matter of practice it is never worth calculating a correlation coefficient for so few observations: the figures are given solely as a short example on which the student can test his knowledge of the work.]

11.2. (Data from W. Little: Labour Commission Report, Vol. 5, Part 1, 1894, and Official Returns.)

The following figures show (1) the estimated average earnings of agricultural labourers, X, (2) the percentage of population in receipt of poor law relief, Y, (3) the ratio of the number of paupers receiving outdoor relief to the number receiving relief in workhouses, Z, for certain districts in England and Wales in 1893.

Find the correlations between X and Y, Y and Z, and Z and X. Draw scatter diagrams to illustrate the various joint distributions.

Union.	Estimated Average Earnings of Agricultural Labourers. Shillings and Pence per Week.		Percentage of Population in Receipt of Poor Law Relief.	Ratio of Number of Paupers Receiving Outdoor Relief to the Number Receiving Relief in Workhouses.
	s.	d.		
1. Glendale	20	9	2·40	6·40
2. Wigton	20	3	2·29	4·04
3. Garstang	19	8	1·39	7·90
4. Belper	18	6	1·92	3·31
5. Nantwich	17	8	2·98	7·85
6. Atcham	17	6	1·17	0·45
7. Driffield	17	1	3·79	10·00
8. Uttoxeter	17	0	3·01	4·43
9. Wetherby	17	0	2·39	4·78
10. Easingwold	16	11	2·78	4·73
11. Southwell	16	6	3·09	6·66
12. Hollingbourn	16	4	2·78	1·22
13. Melton Mowbray	16	3	2·61	4·27
14. Truro	16	3	4·33	7·50
15. Godstone	16	0	3·02	4·44
16. Louth	16	0	4·20	8·34
17. Brixworth	15	9	1·29	0·69
18. Crediton	15	8	5·16	9·89
19. Holbeach	15	6	4·75	4·00
20. Maldon	15	6	4·64	6·02
21. Monmouth	15	4	4·26	8·27
22. St. Neots	15	3	1·66	1·58
23. Swaffham	15	0	5·37	16·04
24. Thakeham	15	0	3·38	1·96
25. Thame	15	0	5·84	9·28
26. Thingoe	15	0	4·63	8·72
27. Basingstoke	15	0	3·93	2·97
28. Cirencester	15	0	4·54	5·38
29. North Witchford	14	10	3·42	3·24
30. Pewsey	14	9	5·88	7·61
31. Bromyard	14	9	4·36	5·87
32. Wantage	14	9	3·85	5·50
33. Stratford-on-Avon	14	7	3·92	3·58
34. Dorchester	14	6	4·48	6·93
35. Woburn	14	6	5·67	6·02
36. Buntingford	14	4	4·91	4·92
37. Pershore	13	6	4·34	4·64
38. Langport	12	6	5·19	10·56

11.3. Verify the following data for the under-mentioned tables of this chapter. Calculate the means of rows and columns and draw a diagram showing the lines of regression for the data of Table 11.1. (Sheppard's correction used only in Table 11.4.)

	11.1.	11.3.	11.4.	11.6.
Mean of X . . .	55·3 mm.	67·70 in.	6·22 years	509·2
,, ,, Y . . .	53·1 ,,	68·66 ,,	18·61 galls.	14,500
Standard deviation of X .	6·86 ,,	2·72 ,,	2·21 years	7·46
,, ,, Y .	5·77 ,,	2·75 ,,	3·37 galls.	18,100
Coefficient of correlation .	+0·97	+0·51	+0·22	−0·014
Coefficient of contingency (for the grouping stated below).	0·90	0·51	0·26	0·47

In calculating the coefficient of contingency (coefficient of mean square contingency) use the following groupings, so as to avoid small scattered frequencies at the extremities of the tables and also excessive arithmetic:—

Table 11.1. Group together (1) two top rows, (2) three bottom rows, (3) two first columns, (4) four last columns, leaving centre of table as it stands.

Table 11.3. Regroup by 2-inch intervals, 58·5–60·5, etc., for father, 59·5–61·5, etc., for son. If a 3-inch grouping be used (58·5–61·5, etc., for both father and son), the coefficient of mean square contingency is 0·465. (Both results cited from Pearson, ref. (84).)

Table 11.4. For columns, group those headed 3 and 4, 5 and 6, 7 and 8, 9 and 10, 11 and over; for rows, group those headed 8–11, 12–13, 14–15, 16–17, 18–19, 20–21, 22–23, 24–25, 26–27, 28 and over.

Table 11.6. For columns, group all up to 494·5 and all over 521·5, leaving central columns. Rows, singly up to 20: then 20–28, 28–44, 44–56, 56 upwards.

11.4. (Data from Statistical Review of England and Wales for 1933, Tables, Part 1, p. 3, and Part 2, p. 6.) The following show mean annual birth and death rates in England and Wales for quinquennia since 1876. Find the correlation between birth and death rates.

Period.	Mean Annual Live Birth Rate per 1000 of Population.	Mean Annual Death Rate per 1000 of Population.
1876–80	35·3	20·8
1881–85	33·5	19·4
1886–90	31·4	18·9
1891–95	30·5	18·7
1896–1900	29·3	17·7
1901–1905	28·2	16·0
1906–1910	26·3	14·7
1911–15	23·6	14·3
1916–20	20·1	14·4
1921–25	19·9	12·2
1926–30	16·7	12·1

15

11.5. The following figures (S. Rowson, *Journ. Roy. Stat. Soc.*, vol. 99, 1936) give the relationship between the density of population and seating capacity of cinemas in various districts of Great Britain.

Find the correlation between density of population and proportion of cinemas with (1) seating capacity 500 or less, (2) seating capacity 2000 or more.

District.	Density of Population per square mile.	Percentage of Cinemas.	
		(1) Seating 500 or less.	(2) Seating 2000 or more.
Scotland	163	13·4	4·3
North Wales	165	42·5	0·0
West of England	380	38·2	2·1
Eastern Counties	431	38·8	1·3
South Wales	440	22·4	1·2
North of England	487	16·0	1·2
Yorkshire and district . . .	594	15·5	3·1
Midlands	710	20·2	1·6
Home Counties (excluding London) .	794	28·2	3·0
Lancashire	2157	13·5	3·6

11.6. Show that the coefficient of correlation is the geometric mean of the coefficients of regression; verify from the data of Examples 11.1, 11.2 and 11.3 that the arithmetic mean of the coefficients of regression is greater than the coefficient of correlation.

11.7. The tangent of the difference of angles A and B is given by

$$\tan (A - B) = \frac{\tan A - \tan B}{1 + \tan A \tan B}$$

Deduce that the smaller angle between regression lines is θ, given by

$$\tan \theta = \frac{1 - r^2}{r} \frac{\sigma_x \sigma_y}{\sigma_x{}^2 + \sigma_y{}^2}$$

and interpret this result when $r = 0$ and $r = \pm 1$.

CHAPTER 12.

NORMAL CORRELATION.

The Bivariate Normal Surface.

12.1. Our study of the normal curve in Chapter 10 may be extended to yield a corresponding expression for the frequency-distribution of pairs of values of two variates. This bivariate normal distribution, known also as "**the bivariate normal surface**," "**the normal correlation surface**" or simply "**the normal surface**," occupies a central position in the theory of bivariate frequency-distributions, and bears to them a relation similar to that borne by the normal curve to the frequency-distributions of a single variate.

The normal surface is of great historical importance, as the earlier work on correlation is, almost without exception, based on the assumption of such a distribution ; though when it was recognised that the properties of the correlation coefficient could be deduced, as in Chapter 11, without reference to the form of the distribution of frequency, a knowledge of this special type of frequency-surface ceased to be so essential. But the generalised normal law is of importance in the theory of sampling : it serves to describe very approximately certain actual distributions (*e.g.* of measurements on man) ; and if it can be assumed to hold good, some of the expressions in the theory of correlation, notably the standard deviations of arrays (and, if more than two variables are involved, the partial correlation coefficients), can be assigned more simple and definite meanings than in the general case. The student should, therefore, be familiar with the more fundamental properties of the distribution.

12.2. Consider first the case in which the two variables are completely independent. Let the distributions of frequency for the two variables x_1 and x_2 singly be given by

$$\left. \begin{aligned} y_1 &= y_1' e^{-\frac{x_1^2}{2\sigma_1^2}} \\ y_2 &= y_2' e^{-\frac{x_2^2}{2\sigma_2^2}} \end{aligned} \right\} \qquad . \qquad . \qquad . \qquad . \quad (12.1)$$

Then, assuming independence, the frequency-distribution of pairs of values must, by the rule of independence, be given by

$$y_{12} = y_{12}' e^{-\frac{1}{2}\left(\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2}\right)} \qquad . \qquad . \qquad . \qquad . \quad (12.2)$$

where

$$y_{12}' = \frac{y_1' y_2'}{N} = \frac{N}{2\pi\sigma_1\sigma_2} \qquad . \qquad . \qquad . \qquad . \quad (12.3)$$

227

Equation (12.2) gives a normal correlation surface for one special case, the correlation coefficient being zero. If we put $x_3 =$ a constant, we see that every section of the surface by a vertical plane parallel to the x_1-axis, *i.e.* the distribution of any array of x_1's, is a normal distribution, with the same mean and standard deviation as the total distribution of x_1's; and a similar statement holds for the arrays of x_2's; these properties must hold good, of course, as the two variables are assumed independent (*cf.* 5.18). The contour lines of the surface, that is to say, lines drawn on the surface at a constant height, are a series of similar ellipses with major and minor axes parallel to the axes of x_1 and x_2 and proportional to σ_1 and σ_2, the equations to the contour lines being of the general form

$$\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} = c^2 \qquad . \qquad . \qquad . \qquad . \quad (12.4)$$

Pairs of values of x_1 and x_2 related by an equation of this form are, therefore, equally frequent.

12.3. Now suppose we have two correlated variates x_1 and x_2, and let the regression of x_1 on x_2 be b_{12} and that of x_2 on x_1 be b_{21}. Let r_{12} be the coefficient of correlation between x_1 and x_2.

Consider the new variates defined by the equations

$$x_{1.2} = x_1 - b_{12}x_2$$
$$x_{2.1} = x_2 - b_{21}x_1$$

This is a notation which we shall later extend considerably.

Then x_1 and $x_{2.1}$ are uncorrelated, as are x_2 and $x_{1.2}$.

For

$$S(x_1 x_{2.1}) = S\{x_1(x_2 - b_{21}x_1)\}$$
$$= S(x_1 x_2) - b_{21}S(x_1)^2$$
$$\frac{1}{N}S(x_1 x_{2.1}) = r_{12}\sigma_{x_1}\sigma_{x_2} - \frac{r_{12}\sigma_{x_2}}{\sigma_{x_1}}\sigma_{x_1}^2$$
$$= 0$$

and similarly for $S(x_2 x_{1.2})$.

Writing σ_1, σ_2 for the standard deviations of x_1, x_2, we see that the standard deviation $\sigma_{1.2}$ of $x_{1.2}$ is given by

$$\sigma_{1.2}^2 = \frac{1}{N}S(x_{1.2}^2) = \frac{1}{N}S(x_1 - b_{12}x_2)^2$$
$$= \{\sigma_1^2 - 2b_{12}r_{12}\sigma_1\sigma_2 + b_{12}^2\sigma_2^2\}$$
$$= \{\sigma_1^2 - 2r_{12}^2\sigma_1^2 + r_{12}^2\sigma_1^2\}$$
$$= \sigma_1^2(1 - r_{12}^2)$$

and similarly $\sigma_{2.1}$ the standard deviation of $x_{2.1}$ is given by

$$\sigma_{2.1}^2 = \sigma_2^2(1 - r_{12}^2)$$

We obtained these results in a slightly different form in **11.22 and 11.24.**

12.4. Suppose further that x_1 and $x_{2.1}$ are not only uncorrelated, but independent, and that each is normally distributed.

In accordance with equation (12.2), we must have for the frequency-distribution of pairs of deviations of x_1 and $x_{2.1}$

$$y_{12} = y_{12}' e^{-\frac{1}{2}\left(\frac{x_1^2}{\sigma_1^2} + \frac{x_{2.1}^2}{\sigma_{2.1}^2}\right)} \qquad . \qquad . \qquad . \qquad (12.5)$$

But

$$\frac{x_1^2}{\sigma_1^2} + \frac{x_{2.1}^2}{\sigma_{2.1}^2} = \frac{x_1^2}{\sigma_1^2(1-r_{12}^2)} + \frac{x_2^2}{\sigma_2^2(1-r_{12}^2)} - 2r_{12}\frac{x_1 x_2}{\sigma_1 \sigma_2(1-r_{12}^2)}$$

$$= \frac{x_1^2}{\sigma_{1.2}^2} + \frac{x_2^2}{\sigma_{2.1}^2} - 2r_{12}\frac{x_1 x_2}{\sigma_{1.2}\sigma_{2.1}}$$

Evidently we should also have arrived at precisely the same expression if we had taken the distribution of frequency for x_2 and $x_{1.2}$, and reduced the exponent

$$\frac{x_2^2}{\sigma_2^2} + \frac{x_{1.2}^2}{\sigma_{1.2}^2}$$

We have, therefore, the general expression for the normal correlation surface for two variables:

$$y_{12} = y_{12}' e^{-\frac{1}{2}\left(\frac{x_1^2}{\sigma_{1.2}^2} + \frac{x_2^2}{\sigma_{2.1}^2} - 2r_{12}\frac{x_1 x_2}{\sigma_{1.2}\sigma_{2.1}}\right)} \qquad . \qquad . \qquad (12.6)$$

Further, since x_1 and $x_{2.1}$, x_2 and $x_{1.2}$, are independent, we must have:

$$y_{12}' = \frac{N}{2\pi\sigma_1\sigma_{2.1}} = \frac{N}{2\pi\sigma_2\sigma_{1.2}} = \frac{N}{2\pi\sigma_1\sigma_2(1-r_{12}^2)^{\frac{1}{2}}} \qquad . \qquad . \qquad (12.7)$$

Expressing $\sigma_{1.2}$ and $\sigma_{2.1}$ in terms of σ_1, σ_2 and r_{12}, we have the alternative form

$$y_{12} = \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r_{12}^2}} e^{-\frac{1}{2(1-r_{12}^2)}\left\{\frac{x_1^2}{\sigma_1^2} - \frac{2r_{12}x_1 x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\right\}} \qquad . \qquad . \qquad (12.8)$$

Properties of the Normal Surface.

12.5. For any given value h_2 of x_2 the distribution of the array of x_1's is given by

$$y_{12} = y_{12}' e^{-\frac{1}{2}\left(\frac{x_1^2}{\sigma_{1.2}^2} + \frac{h_2^2}{\sigma_{2.1}^2} - 2r_{12}\frac{x_1 h_2}{\sigma_{1.2}\sigma_{2.1}}\right)}$$

$$= y_{12}' e^{-\frac{h_2^2}{2\sigma^2} - \frac{\left(x_1 - r_{12}\frac{\sigma_1}{\sigma_2}h_2\right)^2}{2\sigma_{1.2}^2}}$$

This is a normal distribution of standard deviation $\sigma_{1.2}$, with a mean deviating by $r_{12}\frac{\sigma_1}{\sigma_2}h_2$ from the mean of the whole distribution of x_1's.

Hence, since h_2 may be any value, we have the important results :

    (1) that the standard deviations of all arrays of x_1 are the same, and equal to $\sigma_{1.2}$;

    (2) that the regression of x_1 on x_2 is strictly linear.

Similarly, it follows that the s.d.'s of all arrays of x_2 are equal to $\sigma_{2.1}$, and that the regression of x_2 on x_1 is linear.

    12.6. The contour lines are, as in the case of independence, a series of concentric and similar ellipses ; the major and minor axes are, however,
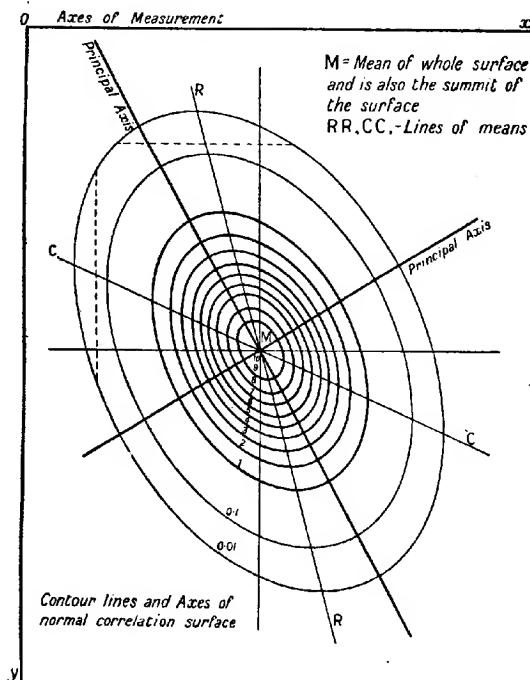
FIG. 12.1.—Principal Axes and Contour Lines of the Normal Correlation Surface.

no longer parallel to the axes of x_1 and x_2, but make a certain angle with them. Fig. 12.1 illustrates the calculated form of the contour lines for one case, RR and CC being the lines of regression. As each line of regression cuts every array of x_1 or of x_2 in its mean, and as the distribution

of every array is symmetrical about its mean, RR must bisect every horizontal chord and CC every vertical chord, as illustrated by the two chords shown by dotted lines ; it also follows that RR cuts all the ellipses in the points of contact of the horizontal tangents to the ellipses, and CC in the points of contact of the vertical tangents. The surface or solid itself, somewhat truncated, is shown in fig. 11.1, page 204.

12.7. Since, as we see from fig. 12.1, a normal surface for two correlated variables may be regarded merely as a certain surface for which r is zero turned round through some angle, and since for every angle through which it is turned the distributions of all x_1 arrays and x_2 arrays are normal, it follows that every section of a normal surface by a vertical plane is a normal curve, *i.e.* the distributions of arrays taken at any angle across the surface are normal.

12.8. It also follows that, since the total distributions of x_1 and x_2 must be normal for every angle through which the surface is turned, the distributions of totals given by slices or arrays taken at any angle across a normal surface must be normal distributions. But these would give the distributions of functions like $ax_1 \pm bx_2$, and consequently (1) the distribution of any linear function of two normally distributed variables x_1 and x_2 must also be normal ; (2) the correlation between any two linear functions of two normally distributed variables must be normal correlation.

Result (1) is very important, and may easily be extended to cover the case of n variables $x_1 \ldots x_n$. Suppose, in fact, we have n such variables each of which is normally distributed, and a linear function $ax_1 + bx_2 + \ldots + hx_n$. Since $ax_1 + bx_2$ is normally distributed, $(ax_1 + bx_2) + cx_3$ is normally distributed, and hence so is $(ax_1 + bx_2 + cx_3) + dx_4$, and so on. Thus the function $ax_1 + \ldots + hx_n$ is normally distributed.

Hence, the sum of n normal variates is distributed normally ; and in particular the mean of n normal variates is distributed normally. More particularly still, the mean of samples of n from a normal universe is normally distributed.

12.9. Returning to the normal surface, it is interesting to inquire what is the angle θ through which the surface has been turned from the position for which the correlation was zero. The major and minor axes of the ellipses are sometimes termed the **principal axes.** If ξ_1, ξ_2 be the co-ordinates referred to the principal axes (the ξ_1-axis being the x_1-axis in its new position), we have for the relation between ξ_1, ξ_2, x_1, x_2, the angle θ being taken as positive for a rotation of the x_1-axis which will make it, if continued through 90°, coincide in direction and sense with the x_2-axis,

$$\left. \begin{array}{l} \xi_1 = x_1 \cos \theta + x_2 \sin \theta \\ \xi_2 = x_2 \cos \theta - x_1 \sin \theta \end{array} \right\} \qquad . \qquad . \qquad (12.9)$$

But, since ξ_1, ξ_2 are uncorrelated, $S(\xi_1 \xi_2) = 0$. Hence, multiplying together equations (12.9) and summing.

$$0 = (\sigma_2{}^2 - \sigma_1{}^2) \sin 2\theta + 2r_{12}\sigma_1\sigma_2 \cos 2\theta$$

$$\tan 2\theta = \frac{2r_{12}\sigma_1\sigma_2}{\sigma_1{}^2 - \sigma_2{}^2} \qquad . \qquad . \qquad . \qquad (12.10)$$

It should be noticed that if we *define* the principal axes of any distribution for two variables as being a pair of axes at right angles for which the

variables ξ_1, ξ_2 are uncorrelated, equation (12.10) gives the angle that they make with the axes of measurement whether the distribution be normal or not.

12.10. The two standard deviations, say S_1 and S_2, about the principal axes are of some interest, for evidently from **12.2** the major and minor axes of the contour ellipses are proportional to these two standard deviations. They may be most readily determined as follows. Squaring the two transformation equations (12.9), summing and adding, we have:

$$S_1{}^2 + S_2{}^2 = \sigma_1{}^2 + \sigma_2{}^2 \quad . \qquad . \qquad . \quad (12.11)$$

Referring the surface to the axes of measurement, we have for the central ordinate, by equation (12.7),

$$y'_{12} = \frac{N}{2\pi\sigma_1\sigma_2(1-r_{12}^2)^{\frac{1}{2}}}$$

Referring it to the principal axes, by equation (12.3),

$$y'_{12} = \frac{N}{2\pi S_1 S_2}$$

But these two values of the central ordinate must be equal, therefore

$$S_1 S_2 = \sigma_1\sigma_2(1-r_{12}^2)^{\frac{1}{2}} \qquad . \qquad . \qquad . \quad (12.12)$$

(12.11) and (12.12) are a pair of simultaneous equations from which S_1 and S_2 may be very simply obtained in any arithmetical case. Care must, however, be taken to give the correct signs to the square root in solving. $S_1 + S_2$ is necessarily positive, and $S_1 - S_2$ also if r is positive, the major axes of the ellipses lying along ξ_1; but if r be negative, $S_1 - S_2$ is also negative. It should be noted that, while we have deduced (12.12) from a simple consideration depending on the normality of the distribution, it is really of general application (like equation (12.11)), and may be obtained at somewhat greater length from the equations for transforming co-ordinates.

12.11. As an example of the application of the foregoing theory to a practical case, we proceed to consider the distribution of Table 11.3, page 199, showing the correlation between stature of father and son, and to test, as far as we can by elementary methods, whether a normal surface will fit the data.

12.12. The first important property of the normal distribution is the linearity of regression. This was well illustrated for these data in fig. 11.8 (p. 211). Subject to some investigation as to the deviations from strict linearity which may occur as the result of sampling fluctuations, we may conclude that the regression is appreciably linear. We shall consider a test of linearity in later chapters (see Chapter 23).

12.13. The second important property is the constancy of the standard deviation for all parallel arrays.

The standard deviations of the ten columns from that headed 62·5–63·5 onwards are :

2·56	2·60
2·11	2·26
2·55	2·26
2·24	2·45
2·23	2·38

the mean being 2·36. The standard deviations again only fluctuate irregularly round their mean value. The mean of the first five is 2·34, of the second five 2·38, a difference of only 0·04 ; of the first group, two are greater and three are less than the mean, and the same is true of the second group. There does not seem to be any indication of a general tendency for the standard deviation to increase or decrease as we pass from one end of the table to the other. We are not yet in a position to test how far the differences from the average standard deviation might have arisen in sampling from a record in which the distribution was strictly normal, but, as a fact, a rough test suggests that they might have done so.

12.14. Next we note that the distributions of all arrays of a normal surface should themselves be normal. Owing, however, to the small numbers of observations in any array, the distributions of arrays are very irregular, and their normality cannot be tested in any very satisfactory way ; we can only say that they do not exhibit any marked or regular asymmetry. But we can test the allied property of a normal correlation table, viz. that the totals of arrays must give a normal distribution even if the arrays be taken diagonally across the surface, and not parallel to either axis of measurement. From an ordinary correlation table we cannot find the totals of such diagonal arrays exactly, but the totals of arrays at an angle of 45° will be given with sufficient accuracy for our present purpose by the totals of lines of diagonally adjacent compartments. Referring again to Table 11.3, and forming the totals of such diagonals (running up from left to right), we find, starting at the top left-hand corner of the table, the following distribution :—

0·25	78·75
2	81·25
3·25	66·5
6·25	59·25
8	42·25
9·75	30·75
17	29·25
34·5	19
42	10·75
46·25	7
60·5	4·25
67·5	3·5
85·75	1·75
87·25	1
78	0·25
94·25	

Total   1078

The mean of this distribution is at 0·359 of an interval above the centre of the interval with frequency 78 ; its standard deviation is 4·757 intervals, or, remembering that the interval is $1/\sqrt{2}$ of an inch, 3·364 inches. (This value may be checked directly from the constants for the table given in Exercise 11.3, page 225, for we have, from the first of the transformation equations (12.9),

$$\sigma_\xi{}^2 = \sigma_1{}^2 \cos^2 \theta + \sigma_2{}^2 \sin^2 \theta + 2r_{12}\sigma_1\sigma_2 \sin \theta \cos \theta$$

and inserting $\sigma_1 = 2\cdot72$, $\sigma_2 = 2\cdot75$, $r_{12} = 0\cdot51$, $\sin\theta = \cos\theta = 1/\sqrt{2}$, find $\sigma_\xi = 3\cdot361$.) Drawing a diagram and fitting a normal curve, we have fig. 12.2; the distribution is rather irregular but the fit is fair; certainly there is no marked asymmetry, and, so far as the graphical test goes, the distribution may be regarded as appreciably normal. One of the greatest divergences of the actual distribution from the normal curve occurs in the almost central interval with frequency 78; the difference between the observed and calculated frequencies is here 12 units, but nevertheless it
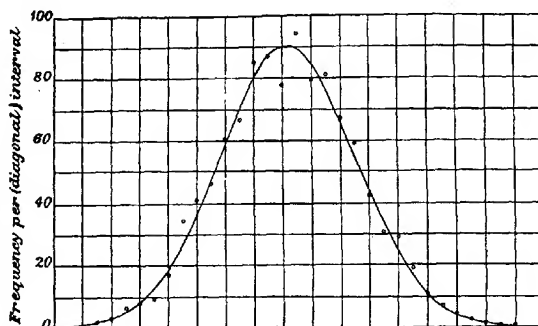
FIG. 12.2.—Distribution of Frequency obtained by Addition of Table 11.3 along Diagonals running up from left to right, fitted with a Normal Curve.

may well have occurred as a fluctuation of sampling. In fact, anticipating our discussion of the use of the standard error (standard deviation of simple sampling) in testing the significance of sampling fluctuations (19.4), we may note that the standard error in this case is $\sqrt{npq}$, where n is the number of observations and p and q the chances of an individual falling or not falling within the given interval. p may be taken as $90/1078$, and therefore the standard error is

$$\sqrt{1078 \cdot \frac{90}{1078} \cdot \frac{988}{1078}} = 9\cdot1$$

The observed deviation, 12, is not much greater than this and may therefore have occurred as a sampling fluctuation. We have used here the exact expression for the standard error, but since p is small we might have used the approximation $\sqrt{pn} = \sqrt{90} = 9\cdot5$. This last is useful as giving a test which can be applied on sight.

12.15. So far, we have seen (1) that the regression is approximately linear; (2) that, in the arrays which we have tested, the standard deviations are approximately constant, or at least that their differences are only small, irregular and fluctuating; (3) that the distribution of totals for one set of diagonal arrays is approximately normal. These results suggest, though they cannot completely prove, that the whole distribution of frequency may be regarded as approximately normal,

within the limits of fluctuations of sampling. We may therefore apply a
more searching test, viz. the form of the contour lines and the closeness
of their fit to the contour ellipses of the normal surface. It may, however,
be seen that no very close fit can be expected. Since the frequencies in
the compartments of the table are small, the standard error of any
frequency is given approximately by its square root (19.15), and this
implies a standard error of about 5 units at the centre of the table, 3 units
for a frequency of 9, or 2 units for a frequency of 4 : fluctuations of these
magnitudes are quite possible and might cause wide divergences in the
corresponding contour lines.

12.16. Using the suffix 1 to denote the constants relating to the
distribution of stature for fathers, and 2 the same constants for the sons,

$$N = 1078 \quad M_1 = 67.70 \quad M_2 = 68.66 \quad r_{12} = 0.51$$
$$\sigma_1 = 2.72 \quad \sigma_2 = 2.75$$

Hence we have from equation (12.7),

$$y_{12}' = 26.7$$

and the complete expression for the fitted normal surface is

$$y = 26.7 e^{-\frac{1}{2}\left(\frac{x_1^2}{5.47} + \frac{x_2^2}{5.60} - \frac{x_1 x_2}{5.43}\right)}$$

The equation to any contour ellipse will be given by equating the index
of e to a constant, but it is very much easier to draw the ellipses if we refer
them to their principal axes. To do this we must first determine θ, S_1
and S_2. From (12.10),

$$\tan 2\theta = -46.49$$

whence $2\theta = 91° 14'$, $\theta = 45° 37'$, the principal axes standing very nearly
at an angle of 45° with the axes of measurement, owing to the two standard
deviations being very nearly equal. They should be set off on the diagram,
not with a protractor, but by taking $\tan \theta$ from the tables (1.022) and
calculating points on each axis on either side of the mean.

To obtain S_1 and S_2 we have, from (12.11) and (12.12),

$$S_1{}^2 + S_2{}^2 = 14.961$$
$$2 S_1 S_2 = 12.868$$

Adding and subtracting these equations from each other and taking the
square root,

$$S_1 + S_2 = 5.275$$
$$S_1 - S_2 = 1.447$$

whence $S_1 = 3.36$, $S_2 = 1.91$; owing to the principal axes standing nearly
at 45° the first value is sensibly the same as that found for σ_ξ in 12.14.
The equations to the contour ellipses, referred to the principal axes, may
therefore be written in the form

$$\frac{\xi_1{}^2}{(3.36)^2} + \frac{\xi_2{}^2}{(1.91)^2} = c^2$$

the major and minor semi-axes being $3.36 \times c$ and $1.91 \times c$ respectively. To find c for any assigned value of the frequency y we have:

$$y_{12} = y'_{12} e^{-\frac{1}{2}c^2}$$
$$c^2 = \frac{2(\log y'_{12} - \log y_{12})}{\log e}$$

Supposing that we desire to draw the three contour ellipses for $y = 5$, 10 and 20, we find $c = 1.83$, 1·40 and 0·76, or the following values for the major and minor axes of the ellipses : semi-major axes, 6·15, 4·70, 2·55 ; semi-minor axes, 3·50, 2·67, 1·45. The ellipses drawn with these axes are shown in fig. 12.3, very much reduced, of course, from the original
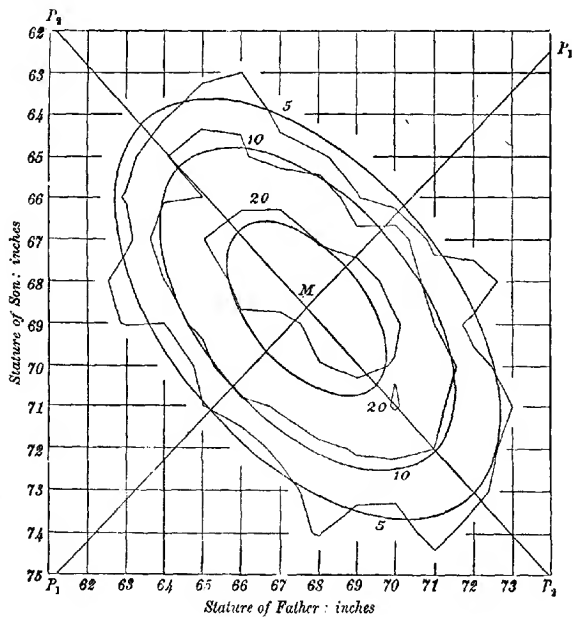
Fɪɢ. 12.3.—Contour Lines for the Frequencies 5, 10 and 20 of the Distribution of Table 11.3, and corresponding Contour Ellipses of the Fitted Normal Surface. P_1P_1, P_2P_2, principal axes; M, mean.

drawing, one of the squares shown representing a square inch on the original. The actual contour lines for the same frequencies are shown by the irregular polygons superposed on the ellipses, the points on these polygons having been obtained by simple graphical interpolation between the frequencies in each row and each column—diagonal interpolation between the frequencies in a row and the frequencies in a column not being used. It will be seen that the fit of the two lower contours is, on

the whole, fair, especially considering the high standard errors. In the case of the central contour, $y = 20$, the fit looks very poor to the eye, but if the ellipse be compared carefully with the table, the figures suggest that here again we have only to deal with the effects of fluctuations of sampling. For father's stature $= 66$ in., son's stature $= 70$ in., there is a frequency of 18·75, and an increase in this much less than the standard error would bring the actual contour outside the ellipse. Again, for father's stature $= 68$ in., son's stature $= 71$ in., there is a frequency of 19, and an increase of a single unit would give a point on the actual contour below the ellipse. Taking the results as a whole, the fit must be considered quite as good as we could expect with such small frequencies.

It is perhaps of historical interest to note that Sir Francis Galton, working without a knowledge of the theory of normal correlation, suggested that the contour lines of a similar table for the inheritance of stature seemed to be closely represented by a series of concentric and similar ellipses (ref. (250)) : the suggestion was confirmed when he handed the problem, in abstract terms, to a mathematician, J. D. Hamilton Dickson (ref. (252)), asking him to investigate " the Surface of Frequency of Error that would result from these data, and the various shapes and other particulars of its sections that were made by horizontal planes."

Isotropic Character of the Normal Surface.

12.17. The normal distribution of frequency for two variables is an isotropic distribution, to which all the theorems of **5.16** apply. For if we isolate the four compartments of the correlation table common to the rows and columns centring round values of the variables x_1, x_2, x_1', x_2', we have for the ratio of the cross-products (frequency of $x_1 x_2$ multiplied by frequency of $x_1' x_2'$, divided by frequency of $x_1 x_2'$ multiplied by frequency of $x_1' x_2$),

$$e^{\frac{r_{12}}{\sigma_{1.2}\sigma_{2.1}}(x_1' - x_1)(x_2' - x_2)}$$

Assuming that $x_1' - x_1$ has been taken of the same sign as $x_2' - x_2$, the exponent is of the same sign as r_{12}. Hence, the association for this group of four frequencies is also of the same sign as r_{12}, the ratio of the cross-products being unity, or the association zero, if r_{12} is zero. In a normal distribution, the association is therefore of the same sign—the sign of r_{12}—for every tetrad of frequencies in the compartments common to two rows and two columns ; that is to say, the distribution is isotropic. It follows that every grouping of a normal distribution is isotropic whether the class-intervals are equal or unequal, large or small, and the sign of the association for a normal distribution grouped down to 2×2-fold form must always be the same whatever the axes of division chosen.

12.18. These theorems are of importance in the applications of the theory of normal correlation to the treatment of qualitative characters which are subjected to a manifold classification. The contingency tables for such characters are sometimes regarded as groupings of a normal distribution of frequency, and the coefficient of correlation is determined on this hypothesis by a rather lengthy procedure (see below, **13.23**, page 251). Before applying this procedure it is well, therefore, to see

whether the distribution of frequency may be regarded as approximately isotropic, or reducible to isotropic form by some alteration in the order of rows and columns (**5.16** and **5.17**). If only reducible to isotropic form by some rearrangement, this rearrangement should be effected before grouping the table to 2- × 2-fold form for the calculation of the correlation coefficient by the process referred to. If the table is not reducible to isotropic form by any rearrangement, the process of calculating the coefficient of correlation on the assumption of normality is to be avoided. Clearly, even if the table be isotropic it need not be normal, but at least the test for isotropy affords a rapid and simple means for excluding certain distributions which are not even remotely normal. Table 5.2, page 66, might possibly be regarded as a grouping of normally distributed frequency if rearranged as suggested in **5.15**— it would be worth the investigator's while to proceed further and compare the actual distribution with a fitted normal distribution—but Table 5.4 could not be regarded as normal, and could not be rearranged so as to give a grouping of normally distributed frequency.

12.19. If the frequencies in a contingency table be not large, and also if the contingency or correlation be small, the influence of casual irregularities due to fluctuations of sampling may render it difficult to say whether the distribution may be regarded as essentially isotropic or not. In such cases some further condensation of the table by grouping together adjacent rows and columns, or some process of " smoothing " by averaging the frequencies in adjacent compartments, may be of service. The correlation table for stature in father and son (Table 11.3), for instance, is obviously not strictly isotropic as it stands : we have seen, however, that it appears to be normal, within the limits of fluctuations of sampling, and it should consequently be isotropic within such limits. We can apply a rough test by regrouping the table in a much coarser form, say with four rows and four columns : the table below exhibits such a grouping, the limits of rows and of columns having been so fixed as to include not less than 200 observations in each array.

TABLE 12.1.—(Condensed from Table 11.3, p. 199.)

Son's Stature (inches).	Father's Stature (inches).				
	Under 65·5.	65·5–67·5.	67·5–69·5.	69·5 and over.	Total.
Under 66·5	97·5	74·25	34·75	10·5	217
66·5–68·5	76·5	108	85	52	321·5
68·5–70·5	33·25	64·75	95	84·5	277·5
70·5 and over	14·75	32·5	80·75	134	262
Total	222	279·5	295·5	281	1078

Taking the ratio of the frequency in column 1 to the sum of the frequencies in columns 1 and 2 for each successive row, and so on for the other pairs of columns, we find the following series of ratios :—

TABLE 12.2.—*Ratio of Frequency in Column m to Frequency in Column m + Frequency in Column (m + 1) of Table 12.1.*

Row.	Columns		
	1 and 2.	2 and 3.	3 and 4.
1	0·568	0·681	0·768
2	0·415	0·560	0·620
3	0·339	0·405	0·529
4	0·312	0·287	0·376

These ratios decrease continuously as we pass from the top to the bottom of the table, and the distribution, as condensed, is therefore isotropic. The student should form one or two other condensations of the original table to 3- × 3- or 4- × 4-fold form: he will probably find them either isotropic or diverging so slightly from isotropy that an alteration of the frequencies, well within the margin of possible fluctuations of sampling, will render the distribution isotropic.

Relationship between Contingency and Normal Correlation.

12.20. It was shown by Karl Pearson that if a normal bivariate universe is divided into sections so as to form a contingency table, the coefficient of mean square contingency, C, tends to the value r in magnitude as the intervals become finer and finer, though of course it is always positive in sign. It was, in fact, the relation

$$r = \pm \sqrt{\frac{\phi^2}{1+\phi^2}}$$

where ϕ^2 is the mean square contingency, which led Pearson to identify C with the expression on the right.

The values of C and r for the distributions of some of the tables of Chapter 11 were compared in Exercise 11.3, page 225.

SUMMARY.

1. The equation of the normal surface is

$$y_{12} = \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r_{12}^2}} e^{-\frac{1}{2(1-r_{12}^2)}\left\{\frac{x_1^2}{\sigma_1^2} - \frac{2r_{12}x_1x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\right\}}$$

where σ_1 is the s.d. of x_1, σ_2 that of x_2, and r_{12} the correlation between x_1 and x_2.

This may also be written

$$y_{12} = \frac{N\sqrt{1-r^2}}{2\pi\sigma_{1.2}\sigma_{2.1}} e^{-\frac{1}{2}\left\{\frac{x_1^2}{\sigma_{1.2}^2} - \frac{2r_{12}x_1x_2}{\sigma_{1.2}\sigma_{2.1}} + \frac{x_2^2}{\sigma_{2.1}^2}\right\}}$$

where

$$\sigma_{1.2}^2 = \sigma_1^2(1 - r_{12}^2), \qquad \sigma_{2.1}^2 = \sigma_2^2(1 - r_{12}^2)$$

2. For two variates normally correlated the standard deviations of parallel arrays are equal and the regressions are linear.

3. Any section of the normal surface by a vertical plane is a normal curve, and a section by a horizontal plane is an ellipse. The ellipses given by horizontal sections are similar and similarly situated.

4. The bivariate normal distribution is isotropic.

5. A linear function of variates, each of which is normally distributed, is also normally distributed.

EXERCISES.

12.1. Deduce equation (12.12) from the equations for transformation of co-ordinates without assuming the normal distribution. (A proof will be found in ref. (248).)

12.2. Hence show that if the pairs of observed values of x_1 and x_2 are represented by points on a plane, and a straight line drawn through the mean, the sum of the squares of the distances of the points from this line is a minimum if the line is the major principal axis.

12.3. The coefficient of correlation with reference to the principal axes being zero, and with reference to other axes *something*, there must be some pair of axes at right angles for which the correlation is a maximum, *i.e.* is numerically greatest without regard to sign. Show that these axes make an angle of 45° with the principal axes, and that the maximum value of the correlation is

$$\pm \frac{S_1{}^2 - S_2{}^2}{S_1{}^2 + S_2{}^2}$$

12.4. (Sheppard, ref. (258).) A fourfold table is formed from a normal correlation table, taking the points of division between A and a, B and β, at the medians, so that $(A) = (a) = (B) = (\beta) = N/2$. Show that

$$r = \cos\left(1 - \frac{2(AB)}{N}\right)\pi$$

12.5. Show that the points of inflection of the sections of the normal surface by vertical planes through the mean of the distribution lie on an ellipse; and show how this ellipse may be used to give the standard deviations of such sections.

12.6. Hence find the minimum and maximum standard deviations which can be taken by such sections, and show that any specified value of the s.d. between the minimum and maximum will be given by two, and only two, sections.

12.7. Find the conditions that the surface

$$z = ke^{ax^2 + 2hxy + by^2}$$

can represent a normal correlation surface whose variates are x and y. Assuming these conditions satisfied, express σ_1, σ_2 and r_{12} in terms of a, h and b.

FURTHER THEORY OF CORRELATION.

Methods of Estimating the Product-moment Correlation Coefficient.

13.1. The only strict method of calculating the correlation coefficient is that described in Chapter 11, from the formula

$$r = \frac{S(xy)}{\sqrt{S(x^2)S(y^2)}}$$

Where possible this formula should be employed. It sometimes happens, however, owing to incomplete data, that we are constrained to use some method of approximation. Furthermore, the large amount of arithmetical labour involved in applying the ordinary formula may sometimes be avoided by approximations which are sufficiently accurate for the purpose in view. We therefore proceed to give a few methods of this kind. They are not recommended for general use as they will, as a rule, lead to different results in different hands.

13.2. (1) The means of rows and columns are plotted on a diagram, and lines fitted to the points by eye, say by shifting about a stretched black thread until it seems to run as near as may be to all the points. If b_1, b_2 be the slopes of these two lines to the vertical and the horizontal respectively,

$$r = \sqrt{b_1 b_2}$$

Hence the value of r may be estimated from any such diagram as fig. 11.8 or 11.9, in the absence of the original table. Further, if a correlation table be not grouped by equal intervals, it may be difficult to calculate the product sum, but it may still be possible to plot approximately a diagram of the two lines of regression, and so determine roughly the value of r. Similarly, if only the means of two rows and two columns, or of one row and one column in addition to the means of the two variables, are known, it will still be possible to estimate the slopes of RR and CC, and hence the correlation coefficient.

(2) The means of one set of arrays only, say the rows, are calculated, and also the two standard deviations σ_x and σ_y. The means are then plotted on a diagram, using the standard deviation of each variable as the unit of measurement, and a line fitted by eye. The slope of this line to the vertical is r. If the standard deviations be not used as the units of measurement in plotting, the slope of the line to the vertical is $r\sigma_x/\sigma_y$, and hence r will be obtained by dividing the slope by the ratio of the standard deviations.

This method, or some variation of it, is often useful as a makeshift when the data are too incomplete to permit of the proper calculation of the

correlation, only one line of regression and the ratio of the dispersions of the two variables being required : the ratio of the quartile deviations, or other simple measures of dispersion, will serve quite well for rough purposes in lieu of the ratio of standard deviations. As a special case, we may note that if the two dispersions are approximately the same, the slope of RR to the vertical is r.

Plotting the medians of arrays on a diagram with the quartile deviations as units, and measuring the slope of the line,' was the method of determining the correlation coefficient (" Galton's function ") used by Sir Francis Galton, to whom the introduction of such a coefficient is due (refs. (242) and (243), *cf.* also ref. (245)).

(3) If s_x be the standard deviation of errors of estimate like $x - b_1 y$, we have, from **11.24**,

$$s_x{}^2 = \sigma_x{}^2 (1 - r^2)$$

and hence,

$$r = \sqrt{1 - \frac{s_x{}^2}{\sigma_x{}^2}}$$

But if the dispersions of arrays do not differ largely, and the regression is nearly linear, the value of s_x may be estimated from the average of the standard deviations of a few rows, and r determined—or rather estimated —accordingly. Thus in Table 11.3 the standard deviations of the ten columns headed 62·5–63·5, 63·5–64·5, etc., are :

2·56	2·26
2·11	2·26
2·55	2·45
2·24	2·33
2·23	——
2·60	Mean 2·359

The standard deviation of the stature of all sons is 2·75 : hence approximately

$$r = \sqrt{1 - \left(\frac{2·359}{2·75}\right)^2}$$
$$= 0·514$$

This is the same as the value found by the product-sum method to the second decimal place. It would be better to take an average by counting the square of each standard deviation once for each observation in the column (or " weighting " it with the number of observations in the column), but in the present case this would only lead to a very slightly different result, viz. $s = 2·362$, $r = 0·512$.

Non-linear Regression.

13.3. We referred in Chapter 11 to the fact that the treatment of cases when the regression is non-linear is somewhat difficult. We may, by the methods of Chapter 17, and otherwise, fit curves of any order to the means of arrays, just as we have fitted straight lines to them ; but the handling of these regression curves and their interpretation is far more complicated.

13.4. It is therefore desirable, wherever possible, to deal with variates which result in linear regression.   Now it sometimes happens that if a relation between X and Y be suggested, we may, either by theory or by previous experience, throw that relation into the form

$$Y = A + B\phi(X)$$

where A and B are the only unknown constants to be determined.   If a correlation table be then drawn up between Y and $\phi(X)$ instead of Y and X, the regression will be approximately linear.   Thus in Table 11.5, page 201, if X be the rate of discount and Y the percentage of reserves on deposits, a diagram of the curves of regression suggests that the relation between X and Y is approximately of the form

$$X(Y - B) = A$$

A and B being constants ; that is,

$$XY = A + BX$$

Or, if we make XY a new variable, say Z,

$$Z = A + BX$$

Hence, if we draw up a new correlation table between X and Z the regression will probably be much more closely linear.

If the relation between the variables be of the form

$$Y = AB^X$$

we have

$$\log Y = \log A + X \log B$$

and hence the relation between $\log Y$ and X is linear.   Similarly, if the relation be of the form

$$X^n Y = A$$

we have

$$\log Y = \log A - n \log X$$

and so the relation between $\log Y$ and $\log X$ is linear.   By means of such artifices for obtaining correlation tables in which the regression is linear, it may be possible to do a good deal in difficult cases whilst using elementary methods only.   The advanced student should refer to refs. (273) and (377) for different methods of treatment.

The Correlation Ratios.

13.5. In view of the importance of linearity of regression it is desirable to have some criterion which will enable a judgment to be formed whether a regression is, within the limits permitted by sampling fluctuations, linear in any given case.   We now proceed to discuss a coefficient designed for this purpose.

Consider a bivariate frequency table, and let s_{px} be the standard deviation of the pth array of X's.   Let n_p be the number of observations in this array.

Let

$$\sigma_{ax}^2 = \frac{1}{N} S(n_p s_{px}^2) \qquad . \qquad . \qquad . \qquad (13.1)$$

Then a_{xa}^2 is the weighted mean of the variances of arrays, obtained as suggested in the last sentence of **13.2** (3).   Now, let

$$\sigma_{ax}^2 = \sigma_x^2(1 - \eta_{xy}^2) \qquad . \qquad . \qquad . \qquad (13.2)$$

or

$$\eta_{xy}^2 = 1 - \frac{\sigma_{ax}^2}{\sigma_x^2} \qquad . \qquad . \qquad . \qquad (13.3)$$

Then η_{xy} is called the **correlation ratio** of X on Y.   Similarly, η_{yx}, defined by

$$\eta_{yx}^2 = 1 - \frac{\sigma_{ay}^2}{\sigma_y^2}$$

is called the **correlation ratio** of Y on X.

13.6.   The correlation ratios may be put in another form, which is much more convenient for purposes of calculation.

In fact, if M_x is the mean of all the X's and m_{px} the mean of an array, we have, as in equation (8.6),

$$N\sigma_x^2 = S[n_p\{s_{px}^2 + (M_x - m_{px})^2\}]$$

or, using σ_{mx} to denote the standard deviation of m_{px}, obtained by "weighting" each m_{px} according to n_p, the number of observations in the array in which it occurs,

$$\sigma_x^2 = \sigma_{ax}^2 + \sigma_{mx}^2 \qquad . \qquad . \qquad . \qquad (13.4)$$

Hence, substituting in (13.3),

$$\eta_{xy} = \frac{\sigma_{mx}}{\sigma_x} \qquad . \qquad . \qquad . \qquad (13.5)$$

The correlation ratio of X on Y is therefore determined when we have found the standard deviation of X and the standard deviation of the means of its arrays.

13.7.   In **11.22** we saw that

$$\sigma_x^2(1 - r^2) = \frac{1}{N} S(x - b_1 y)^2 \qquad . \qquad . \qquad (13.6)$$

where $x - b_1 y = 0$ is the line of regression of x on y, x and y being the values of X and Y measured from the mean of the distribution.

Now, for any array for which y is constant,

$$\frac{1}{N} S(x - b_1 y)^2 = \frac{1}{N} S\{(x - m_{px}) + (m_{px} - b_1 y)\}^2$$

$$= \frac{n_p}{N} s_{px}^2 + \frac{n_p}{N}(m_{px} - b_1 y)^2$$

the product term vanishing since $S(x - m_{px}) = 0$. Hence, summing for all arrays of y,

$$\sigma_x^2(1 - r^2) = \sigma_{ax}^2 + S\frac{n_p}{N}\left\{(m_{px} - b_1 y)^2\right\}$$

But

$$\sigma_x^2(1 - \eta_{xy}^2) = \sigma_{ax}^2$$

Hence,

$$\sigma_x^2(\eta_{xy}^2 - r^2) = S\left\{\frac{n_p}{N}(m_{px} - b_1 y)^2\right\} \qquad . \qquad . \qquad . \quad (13.7)$$

From this we see that η_{xy} cannot be less than r in absolute value
If $\eta_{xy}^2 = r^2$, then

$$S\{n_p(m_{px} - b_1 y)^2\} = 0$$

i.e.

$$m_{px} - b_1 y = 0$$

for all arrays. This means that the mean m_{px} must be on the line of regression for all arrays, *i.e.* that the regression is linear.

13.8. The divergence of η^2 from r^2 therefore measures the departure of the regression from linearity. It should, however, be noted that sampling fluctuations may cause $\eta^2 - r^2$ to deviate from zero even when the regression is truly linear. We give later a method of testing the significance of observed fluctuations of this kind **(23.44)**.

Calculation of the Correlation Ratio.

13.9. The table on page 246 illustrates the form of the arithmetic for the calculation of the correlation ratio of son's stature on father's stature (Table 11.3). In the first column is given the type of the array (stature of father) ; in the second, the mean stature of sons for that array ; in the third, the difference of the mean of the array from the mean stature of all sons. In the fourth column these differences are squared, and in the sixth they are multiplied by the frequency of the array, two decimal places only having been retained as sufficient for the present purpose. The sum-total of the last column divided by the number of observations (1078) gives $\sigma_{my}^2 = 2.058$, or $\sigma_{my} = 1.43$. As the standard deviation of the sons' stature is 2.75 in., $\eta_{yx} = 0.52$. Before taking the differences for the third column of such a table, it is as well to check the means of the arrays by recalculating from them the mean of the whole distribution, *i.e.* multiplying each array-mean by its frequency, summing and dividing by the number of observations. The form of the arithmetic may be varied, if desired, by working from zero as origin, instead of taking differences from the true mean. The square of the mean must then be subtracted from $S(fm_y^2)/N$ to give σ_{my}^2.

13.10. If the second correlation ratio for this table be worked out in the same way, the value will be found to be the same to the second place of decimals : the two correlation ratios for this table are, therefore, very nearly identical, and only slightly greater than the correlation coefficient (0.51). Both regressions, as follows from the last section, are very nearly linear, a result confirmed by the diagram of the regression lines (fig. 11.8, page 211). On the other hand, it is evident from fig. 11.10, page 213,

EXAMPLE 13.1.—CALCULATION OF THE CORRELATION RATIO: *Son's Stature on Father's Stature : Data of Table 11.3, p. 199.*

1.	2.	3.	4.	5.	6.
Type of Array (Father's Stature).	Mean of Array (Son's Stature).	Difference from Mean of all Sons (68·66).	Square of Difference.	Frequency.	Frequency × (difference)².
59	64·67	− 3·99	15·9201	3	47·76
60	65·64	− 3·02	9·1204	3·5	31·92
61	66·34	− 2·32	5·3824	8	43·06
62	65·56	− 3·10	9·6100	17	163·37
63	66·68	− 1·98	3·9204	33·5	131·33
64	66·74	− 1·92	3·6864	61·5	226·71
65	67·19	− 1·47	2·1609	95·5	206·37
66	67·61	− 1·05	1·1025	142	156·56
67	67·95	− 0·71	0·5041	137·5	69·31
68	69·07	+ 0·41	0·1681	154	25·89
69	69·39	+ 0·73	0·5329	141·5	75·41
70	69·74	+ 1·08	1·1664	116	135·30
71	70·50	+ 1·84	3·3856	78	264·08
72	70·87	+ 2·21	4·8841	49	239·32
73	72·00	+ 3·34	11·1556	28·5	317·93
74	71·50	+ 2·84	8·0656	4	32·26
75	71·73	+ 3·07	9·4249	5·5	51·84
Total	...	...	...	1078	2218·42

$$\sigma^2_{my} = 2218\cdot42/1078 = 2\cdot058 \qquad \sigma_{my} = 1\cdot43$$
$$\eta_{yx} = 1\cdot43/2\cdot75 = 0\cdot52$$

that we should expect the two correlation·ratios for Table 11.6 to differ considerably from each other and from the correlation coefficient. The values found are $\eta_{xy} = 0\cdot14$, $\eta_{yx} = 0\cdot38$ $(r = -0\cdot014)$: η_{xy} is comparatively low as proportions of male births differ little in the successive arrays, but η_{yx} is higher since the line of regression of Y on X is sharply curved. The confirmation of these values is left to the student.

The student should notice that the correlation ratio only affords a satisfactory test when the number of observations is sufficiently large for a grouped correlation table to be formed. In the case of a short series of observations such as that given in Table 11.7, page 203, the method is inapplicable.

The Rank Correlation Coefficient.

13.11. In calculating the coefficient of correlation from the product-moment it is necessary that the data should be definitely measured. If they are not so measured we cannot, in general, determine the coefficient, though we may sometimes approximate to it by one of the methods of **13.2.**

But there may be more serious obstacles than imperfect grouping in the way of finding the correlation between two variates. In the examples

we have considered up to the present the qualities we have discussed have been easily measurable, involving such familiar concepts as height, weight, age and so forth. In certain types of inquiry we may have to deal with qualities which are not expressible as numbers of units of an objective kind.

13.12. Consider, for instance, the relation between mathematical and musical ability in a class of students. " Ability," whether of a general or a specific kind, is a variate in the sense that it varies from one individual to another ; and it may be a numerical variate if we can decide on some unequivocal way of measuring it. A very common mode of attempting to do so is by allotting marks to each student. But such methods are open to many objections, not the least of which is that different examiners would give different marks to the same person. A correlation between the marks obtained for mathematics and music would, therefore, be likely to depend to some extent on the examiner, and would not reflect accurately the relationship between the two qualities.

13.13. Difficulties of this type disappear to some extent if we arrange the students *in order* of their ability, but do not attempt to assess it numerically. There will still be some divergence of opinion between different examiners, perhaps, but it will not as a rule be so serious. We then allot to each student a number which indicates his position in the arrangement according to ability, the first being number 1, the second number 2, and so on. The students are then said to be **ranked**, and the number of a particular individual is his **rank** (*cf.* **8.32**).

13.14. A procedure of this kind is useful in the treatment not only of data which can be ordered but not exactly measured, but of measurable data also. For instance, we can easily rank a number of men according to height without actually measuring them. It is also comparatively easy to rank a number of shades of a colour, or a number of countries according to their importance in the export market, where precise numerical measurement would be very troublesome.

13.15. If we have a set of individuals ranked according to two different qualities it is natural to inquire whether the ranks can be made to give us some measure of the degree of relation between the two qualities.

Suppose we have n individuals, whose ranks according to quality A are $X_1, X_2, X_3, \ldots X_n$, and according to quality B are $Y_1, Y_2, Y_3, \ldots Y_n$, where the X's and Y's are merely permutations of the first n natural numbers. Let $d_k = X_k - Y_k$.

The values of d form a convenient measure of the closeness of the correspondence between A and B. If all the d's are zero the correspondence is perfect, for an individual whose rank is X_k for A will also be X_k for B. We cannot, however, take the sum of the d's as a measure of correspondence, because that sum is zero ; for the sum of the differences of the X's and Y's is the difference of the sums of the X's and the Y's, each of which is the sum of the first n natural numbers.

A possible measure which suggests itself is the sum of the absolute values of the d's, *i.e.* $S|d|$. This measure and its mean $\frac{1}{n} S|d|$ have, in fact, been used, but like the mean deviation **(8.17)** they have certain analytical disadvantages.

13.16. A more convenient coefficient is obtained as follows :—

The values of X range from 1 to n. Their sum is $\dfrac{n(n+1)}{2}$, and their mean is accordingly $\dfrac{n+1}{2}$. This value is also the mean of the Y's.

Let us denote by x_k the value of $X_k - \dfrac{n+1}{2}$, *i.e.* the divergence of X_k from the mean. Similarly for y_k, which we define as $Y_k - \dfrac{n+1}{2}$.

Write

$$\rho = \frac{S(xy)}{\sqrt{S(x^2)S(y^2)}} \qquad . \qquad . \qquad . \qquad (13.8)$$

This is the product-moment coefficient of correlation between X and Y.

We shall call ρ the **rank correlation coefficient.** It may be expressed very simply in terms of n and the d's.

For, as we saw in **8.14**, $S(x^2)$ is $\dfrac{n^3-n}{12}$.

Now,

$$S(d^2) = S(X_k - Y_k)^2 = S(x-y)^2$$
$$= S(x^2) + S(y^2) - 2S(xy)$$

Hence,

$$S(xy) = \tfrac{1}{2}\left\{\frac{n^3-n}{6} - S(d^2)\right\}$$

and substituting in (13.8):

$$\rho = 1 - \frac{6S(d^2)}{n^3 - n} \qquad . \qquad . \qquad . \qquad (13.9)$$

Example 13.2.—The rankings of ten students in mathematics and music are as follows :—

$$\text{Mathematics : } 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$$
$$\text{Music : } 6, 5, 1, 4, 2, 7, 8, 10, 3, 9$$

What is the coefficient of rank correlation ?

The differences d are (mathematical rank minus musical rank)

$$-5, \ -3, \ +2, \ 0, \ +3, \ -1, \ -1, \ -2, \ +6, \ +1$$

These add to zero, as they should.

The squares of d are

$$25, 9, 4, 0, 9, 1, 1, 4, 36, 1$$

which add up to 90.

Hence, from (13.9),

$$\rho = 1 - \frac{540}{990} = +0.45$$

13.17. The rank correlation coefficient varies from $+1$ to -1. If the rank correlation is perfect, all the d's are zero. If, on the other hand, the

ranks are such that the first, second, third in one order correspond to the nth, $(n-1)$th, $(n-2)$th, . . . in the other, $\rho = -1$. The proof is slightly different according to whether n is even or odd. If it is odd, say $= 2m + 1$, the d's are

$$2m, \quad 2m-2, \quad \ldots \quad 2, \quad 0, \quad -2, \quad \ldots \quad -(2m-2), \quad -2m$$

and

$$S(d^2) = 2\{(2m)^2 + (2m-2)^2 + \ldots + 2^2\}$$

$$= \frac{8m(m+1)(2m+1)}{6}$$

Hence,

$$\rho = 1 - \frac{8m(m+1)(2m+1)}{(2m+1)\{(2m+1)^2 - 1\}} = -1$$

If n is even, say $= 2m$,

$$S(d^2) = 2\{(2m-1)^2 + \ldots + 1^2\}$$

$$= \frac{2m}{3}(4m^2 - 1)$$

and

$$\rho = -1 \text{ as before.}[1]$$

Relationship between Rank Correlation and Product-moment Correlation.

13.18. The rank correlation coefficient as we have introduced it is merely a measure, like the coefficients of association, contingency and product-moment correlation, of the correspondence between two quantities. Like those coefficients, it is affected by sampling fluctuations.

It is, however, more easily calculated than most coefficients, and for this reason some writers have advocated its use as a substitute for the product-moment coefficient between the actual measurements, and for estimating the product-moment coefficient from a normal universe. We proceed to examine this practice briefly.

Grade Correlation.

13.19. We referred at the end of Chapter 8 to such quantities as quartiles, deciles and percentiles, which are values of the variate dividing the total frequency into certain specified proportions. For instance, the seventh decile is the variate value such that seven-tenths of the distribution lie below it, *i.e.* exhibit values of the variate less than the decile.

Generally, we may regard the **grade** of an individual as the proportion of individuals which lie below him (*cf.* 8.30). If the universe is continuous, the range of grades will also be continuous.

13.20. To each individual in a bivariate universe there will be attached two grade numbers, one for each variate, and if the universe is

---

[1] The property of varying between $+1$ and -1 does not belong to a similar coefficient proposed by Spearman, and known as his "foot-rule," viz. $R = 1 - \dfrac{3S(|d|)}{n^2 - 1}$.

It may be shown in the above manner that R varies from -0.5 to $+1$, and for this reason alone R seems an undesirable coefficient.

correlated the grades will also be correlated. In fact, Karl Pearson has shown that if the universe is normal, ρ_g, the grade correlation, and r, the ordinary correlation (both calculated by the product-moment method), are related by the equation

$$r = 2 \sin\left(\frac{\pi\rho_g}{6}\right) \qquad . \qquad . \qquad . \qquad . \quad (13.10)$$

13.21. Ranks and grades are connected by a simple relation. In fact, if an individual is of rank k, there are $k - 1$ individuals below him (assuming that the ranking proceeds from the lowest variate value). If we admit, conventionally, that one-half of the individual is to be regarded as lying to the left of the line of division which he makes, and one-half to the right, his grade, g_k, is given by

$$g_k = (k-1) + \tfrac{1}{2} = k - \tfrac{1}{2} \qquad . \qquad . \qquad . \quad (13.11)$$

It follows that *the correlation between ranks is the same as the correlation between grades.* But in a universe which is finite and discontinuous (and ranking is in practice applied to comparatively small universes of twenty or thirty individuals) *it does not follow that*

$$r = 2 \sin\left(\frac{\pi\rho}{6}\right) \qquad . \qquad . \qquad . \qquad . \quad (13.12)$$

Equation (13.10) was obtained by considering grades in a continuous universe, and equation (13.12) is at best an approximation, depending on assumptions which are often of doubtful legitimacy. This is a fact which has not always been appreciated. We may, perhaps, clarify the point by considering the data of Example 13.2.

Example 13.3.—In Example 13.2 we found :

$$\rho = +0.45$$

If we apply (13.12) we find :

$$r = 2 \sin 13.5^{\circ}$$
$$= +0.47$$

Let us consider what this means.

The value r purports to be a correlation coefficient such as would have been obtained by the product-moment method if the two variates had been measurable in the ordinary way. Let us, for the sake of argument, agree that mathematical and musical abilities are capable of measurement.

Now there are only ten members in this universe, and it cannot be regarded with any degree of accuracy as a continuous normal universe. The use of (13.12) in finding the correlation *in the universe of ten* is therefore of doubtful validity, to say the least.

But it is possible to look at this from rather a different point of view, and to regard the ten students as a sample from a practically infinite universe which *is* continuous and normal. The value r is then taken to be an estimate of the correlation coefficient in this universe.

· The legitimacy of this procedure will depend on the extent to which the

grade correlation in the sample can be taken to represent the grade correlation in the universe. It will, we think, be sufficiently evident from the smallness of the sample that the two are likely to diverge considerably owing to sampling fluctuations.

Furthermore, in the comparatively small samples to which (13.12) is applied—the labour of calculating the rank correlation coefficient for large samples is very tedious—it is difficult to obtain any satisfactory evidence from the data themselves that the universe can properly be regarded as normal ; and even if the distribution of each of the variates, taken singly, can be rendered normal by some appropriate transformation of the variate which squeezes or stretches the scale of measurement, it does not necessarily follow that the correlation distribution can in this way be rendered normal.

In practice, moreover, troublesome difficulties sometimes arise owing to two or more individuals being given the same rank. The common procedure of assigning to each individual the average rank of the group, *but nevertheless using* formula (13.9), is inexact.

Use of (13.12) should therefore be made with the utmost reserve. It would probably be better to avoid it altogether and rely on the rank correlation coefficient.

13.22. The relationship between the product-moment coefficient and the rank correlation coefficient might profitably be subjected to further investigation, particularly for small numbers of individuals. As we have just seen, with the present state of our knowledge, the use of the rank coefficient is not to be recommended as a brief method of estimating the product-moment coefficient. It appears, however, to be of service as a quick method of gauging relations between variates which are not normally distributed, or between quantities which cannot readily be measured, when the number of observations is small.

Tetrachoric r.

13.23. To complete our account of methods which have been devised as alternatives to the use of the product-moment correlation coefficient in cases where, for some reason, that coefficient cannot be computed, we may refer to a process specially adapted to the 2 × 2 contingency table.

Consider such a table in the schematic form :

	A	Not-A	Total
B .	a	b	$a+b$
Not-B	c	d	$c+d$
Total	$a+c$	$b+d$	N

Let us assume that our attributes A and B are, in theory, based on measurable quantities ; and let us suppose further that the universe would be normally distributed with respect to those quantities as variates. Then we may regard the above table as the result obtained by dividing a bivariate normal universe into four sections, a division of the X-variate at some point, say h, and a division of the Y-variate at some point k. If we picture the universe as a solid figure, as in fig. 11.1, page 204, the frequencies

a, b, c and d will be the volumes into which the universe is divided by planes perpendicular to the X and Y axes through the points $X = h$ and $Y = k$, respectively.

The problem then arises, given a, b, c and d, what are the values of h and k (in terms of the standard deviations of X and Y), and what is the value of r ?

13.24. A discussion of this problem, which involves some difficult mathematics, is outside the scope of this book. The student may be referred to " *Tables for Statisticians and Biometricians, Parts I and II*," for a short account of the method of solution and for tables which are almost indispensable in working out r for any given case.

A value of r obtained in this way is said to be **tetrachoric.**

The coefficient has often been used to obtain a value of the correlation (so-called) for a contingency table, using some reduction to the four-fold form by amalgamating adjacent arrays, or possibly making more than one such reduction and averaging the results. As such tables are very often far from normal, it is always desirable to test the normality by using more than one reduction. In any case the reader should be informed precisely as to the reduction used.

The Product-moment Correlation Coefficient for a 2 × 2 Table.

13.25. The correlation coefficient is in general only calculated for a table with a considerable number of rows and columns, such as those given in Chapter 11. In some cases, however, a theoretical value is obtainable for the coefficient, which holds good even for the limiting case when there are only two values possible for each variable (*e.g.* 0 and 1) and consequently two rows and two columns (*cf.* Exercises 13.5 and 13.6). It is therefore of some interest to obtain an expression for the coefficient in this case in terms of the class-frequencies.

Using the notation of Chapters 1–4 the table may be written in the form :

Values of Second Variable.	Values of First Variable.		
	X_1	X'_1	Total
X_2	(AB)	(aB)	(B)
X'_2	$(A\beta)$	$(a\beta)$	(β)
Total	(A)	(a)	N

Taking the centre of the table as arbitrary origin and the class-interval, as usual, as the unit, the co-ordinates of the mean are :

$$\bar{\xi} = \frac{1}{2N}\{(a) - (A)\} .$$

$$\bar{\eta} = \frac{1}{2N}\{(\beta) - (B)\}$$

The standard deviations σ_1, σ_2 are given by

$$\sigma_1{}^2 = 0\cdot25 - \bar{\xi}^2 = (A)(a)/N^2$$
$$\sigma_2{}^2 = 0\cdot25 - \bar{\eta}^2 = (B)(\beta)/N^2$$

Finally,
$$S(xy) = \tfrac{1}{4}\{(AB) + (a\beta) - (A\beta) - (aB)\} - N\xi\bar{\eta}$$
Writing
$$(AB) - (A)(B)/N = \delta$$

(as in Chapter 3) and replacing ξ, $\bar{\eta}$ by their values, this reduces to
$$S(xy) = \delta$$
Whence
$$r = \frac{N\delta}{\sqrt{(A)(a)(B)(\beta)}} \qquad . \qquad . \qquad (13.13)$$

This value of r can be used as a coefficient of association, but, unlike the association coefficient of Chapter 3, which is unity if either $(AB) = (A)$ or $(AB) = (B)$, r only becomes unity if $(AB) = (A) = (B)$. This is the only case in which both frequencies (aB) and $(A\beta)$ can vanish so that (AB) and $(a\beta)$ correspond to the frequencies of two points, $X_1 Y_1$, $X_2 Y_2$ on a line. Obviously this alone renders the numerical values of the two coefficients quite incomparable with each other. But further, while the association coefficient is the same for all tables derived from one another by multiplying rows or columns by arbitrary coefficients, the correlation coefficient (13.13) is greatest when $(A) = (a)$ and $(B) = (\beta)$, *i.e.* when the table is symmetrical, and its value is lowered when the symmetrical table is rendered asymmetrical by increasing or reducing the number of A's or B's. For moderate degrees of association, the association coefficient gives much the larger values. The two coefficients possess, in fact, essentially different properties, and are *different* measures of association in the same sense that the geometric and arithmetic means are different forms of average, or the semi-interquartile range and the standard deviation different measures of dispersion.

13.26. The student should realise that the product-sum correlation and the tetrachoric correlation are also two entirely different measures with quite different properties. The one is in no sense an approximation to the other, and the two may often differ largely.

Intraclass Correlation.

13.27. We have previously considered correlations between two distinct types of variate, such as age and yield of milk in cows, or stature of father and stature of son ; but there occurs, mainly in biological studies, a rather different kind of correlation which we will now proceed to discuss.

Suppose we are examining the relationship between the heights of brothers, and consider a pair of brothers. Our two variates will be (1) the height of the first brother, and (2) the height of the second brother. The question is, which are we to regard as the first brother and which as the second ? It is not difficult to lay down rules which would enable us to make a distinction—for instance, we might take the elder brother first, or the taller brother first. But if we did this and drew up a correlation table. for all such pairs, we should not be answering the question as to the relation between brothers in general, for we should only get a correlation between the height of taller brothers and that of shorter brothers, or the height of elder brothers and the height of younger brothers.

13.28. The relationship of brotherhood is in fact symmetrical ; if

A is the brother of *B*, then *B* is the brother of *A*. When we are considering only the relationship in height implied by relationship of blood, there is no relevant character to enable us to single out one brother as the first.

We accordingly treat the problem by taking each pair of brothers in two ways: (1) with the height of *A* as the first variate and that of *B* as the second, and (2) with the height of *B* as the first variate and that of *A* as the second. Similarly, if there are *k* brothers in the family, we enter in the correlation table the results of taking pairs in all possible ways, which number $k(k-1)$. For example, if we have a family containing three brothers with heights 5 ft. 9 in., 5 ft. 10 in. and 5 ft. 11 in., they may be regarded as giving six pairs of variate values :

5 ft. 9 in. with 5 ft. 10 in.	5 ft. 10 in. with 5 ft. 9 in.
5 ft. 9 in. with 5 ft. 11 in.	5 ft. 11 in. with 5 ft. 9 in.
5 ft. 10 in. with 5 ft. 11 in.	5 ft. 11 in. with 5 ft. 10 in.

13.29. Generally, if we have *n* families, each with *k* members, there will be $nk(k-1)$ pairs, and hence the same number of entries in the table.

Such a table is called an **intraclass correlation table,** and the correlation between the two variates is called **intraclass correlation.**

Tables in which all the families have the same number are of particular importance, and we will consider them first. It is, however, permissible to apply the term intraclass correlation to the symmetrical table derived from families which have different numbers of members. This case we shall consider in **13.33.**

13.30. The intraclass correlation table has certain peculiarities, and is not of such a general type as the ordinary table which we have considered hitherto (and which, for the purposes of distinction, is sometimes called an **interclass** table).

Let the variate values in the first family be

$$x_{11}\, x_{12}\, .\, .\, .\, x_{1k}$$

those in the second family being

$$x_{21}\, x_{22}\, .\, .\, .\, x_{2k}$$

and so on, those in the *n*th family being

$$x_{n1}\, x_{n2}\, .\, .\, .\, x_{nk}$$

Consider the mean of the *X*-variate.

In the table the value x_{11} will be associated as an *X*-variate with each of the $(k-1)$ values $x_{12}\, .\, .\, .\, x_{1k}$. Hence it appears $(k-1)$ times. Similarly, every other value appears $(k-1)$ times. Hence the sum of the marginal row, corresponding to the *X*-variate, is $(k-1)S(x)$, the summation extending over all values. But there are $nk(k-1)$ members in the table.

Hence,

$$\bar{X} = \frac{1}{nk(k-1)} (k-1)S(x)$$

$$= \frac{1}{nk} S(x) \qquad . \qquad . \qquad . \qquad . \qquad . \quad (13.14)$$

Similarly,

$$\bar{Y} = \frac{1}{nk} S(x) \qquad . \qquad . \qquad . \qquad . \qquad (13.15)$$

i.e. the means of the variates are the same. This must evidently be the case, for the table is symmetrical.

For the variance of X we have :

$$\sigma_X{}^2 = \frac{1}{nk(k-1)} \{\text{Sum of } (x - \bar{X})^2\}$$

and since each $x - \bar{X}$ occurs $(k-1)$ times,

$$\sigma_X{}^2 = \frac{1}{nk} S(x - \bar{X})^2 . \qquad . \qquad . \qquad . \qquad (13.16)$$

the summation, as before, extending over all the values of x.

Similarly,

$$\sigma_Y{}^2 = \frac{1}{nk} S(x - \bar{Y})^2$$

$$= \frac{1}{nk} S(x - \bar{X})^2$$

$$= \sigma_X{}^2$$

We therefore write

$$\sigma = \sigma_X = \sigma_Y$$

13.31. For the correlation coefficient r we have

$$\sigma^2 r = \frac{1}{nk(k-1)} S'(x_{ij} - \bar{X})(x_{im} - \bar{X}) \qquad . \qquad . \qquad (13.17)$$

where the summation S' extends over all the possible pairs.

We can put this formula into a much simpler form.

Consider the terms in (13.17) for which the first term is $(x_{11} - \bar{X})$. They will be the $(k-1)$ terms of the following series :—

$$(x_{11} - \bar{X})(x_{12} - \bar{X}) + (x_{11} - \bar{X})(x_{13} - \bar{X}) + \ . \ . \ . + (x_{11} - \bar{X})(x_{1k} - \bar{X})$$

$$= (x_{11} - \bar{X})\{(x_{12} + x_{13} + \ . \ . \ . + x_{1k}) - (k-1)\bar{X}\}$$

Now write

$$\bar{X}_1 = \frac{1}{k}(x_{11} + x_{12} + \ . \ . \ . + x_{1k}) . \qquad . \qquad . \qquad (13.18)$$

i.e. $\bar{X}_1$ is the mean of the members of the first family. Then our expression becomes

$$(x_{11} - \bar{X})\{k\bar{X}_1 - x_{11} - (k-1)\bar{X}\}$$

$$= (x_{11} - \bar{X})\{k(\bar{X}_1 - \bar{X}) + \bar{X} - x_{11}\}$$

$$= k(\bar{X}_1 - \bar{X})(x_{11} - \bar{X}) - (x_{11} - \bar{X})^2$$

The sum S' of (13.17) will contain nk such terms.

Hence,

$$nk(k-1)\sigma^2 r = kS(\bar{X}_1 - \bar{X})(x_{11} - \bar{X}) - S(x_{11} - \bar{X})^2 \quad . \quad (13.19)$$

the summation extending over all the nk members.

Now,

$$kS(\bar{X}_1 - \bar{X})(x_{11} - \bar{X})$$

$$=\text{sum of } n \text{ terms like } k \times k(\bar{X}_1 - \bar{X})(\bar{X}_1 - \bar{X})$$

$$=k^2 S''(\bar{X}_1 - \bar{X})^2$$

S'' extending over the n families; and

$$S(x_{11} - \bar{X})^2 = nk\sigma^2$$

Hence, from (13.19),

$$nk(k-1)\sigma^2 r = k^2 S''(\bar{X}_1 - \bar{X})^2 - \sigma^2 nk$$

Now $\dfrac{1}{n} S''(\bar{X}_1 - X)^2$ is the variance of the means of families about the mean of the whole. Calling this σ_m^2, we have

$$nk(k-1)\sigma^2 r = k^2 n\sigma_m^2 - \sigma^2 nk$$

or

$$\{1 + r(k-1)\}\sigma^2 = k\sigma_m^2 \quad . \qquad . \qquad . \quad (13.20)$$

This result gives us the intraclass correlation in terms of the variance of the distribution (according to either variate) and the variance of the means of families.

*Example 13.4.—*In five families of 3 the heights of brothers are : 5' 9", 5' 10", 5' 11" ; 5' 10", 5' 11", 6' 0" ; 5' 11", 6' 0", 6' 1" ; 6' 0", 6' 1", 6' 2" ; 6' 1", 6' 2", 6' 8".   Find the intraclass coefficient of correlation.

Here the mean of the whole $=6'$.

$$\sigma^2 = \frac{1}{5 \times 3}\{9 + 4 + 1 + 4 + 1 + 1 + 1 + 1 + 4 + 1 + 4 + 9\}$$

$$= \frac{40}{15} = \frac{8}{3}$$

$$\sigma_m^2 = \frac{1}{5}\{4 + 1 + 0 + 1 + 4\} = 2$$

Hence, from (13.20),

$$\{1 + 2r\}\frac{8}{3} = 3 \times 2$$

$$1 + 2r = 2 \cdot 25$$

$$r = + 0 \cdot 625$$

13.32.   We may notice two rather unusual results which follow from equation (13.20).

In the first place, since σ_m^2 is not negative,

$$1 + r(k-1) \geqslant 0$$

and hence,

$$r \geqslant -\frac{1}{k-1}$$

Thus, whereas the interclass correlation coefficient can vary from -1 to $+1$, the intraclass coefficient cannot be less than $-\frac{1}{k-1}$. For example, in families of threes the intraclass coefficient cannot be less than $-\frac{1}{2}$.

Secondly, let us consider the correlation within a single family, *i.e.* when $n=1$.

In this case, $\sigma_m^2 = 0$, and hence

$$r = -\frac{1}{k-1}$$

For $k=2$, 3, 4, . . . this gives the successive values of $r=-1$, $-\frac{1}{2}$, $-\frac{1}{3}$, ... It is clear that the first value is correct, for the two values x_1 and x_2 determine only two points $(x_1 x_2)$ and $(x_2 x_1)$, and the slope of the line joining them is negative.

The student should notice that a corresponding negative association will arise between the first and second members of the pair if all possible pairs are chosen from a universe in which the variates can assume only two values, say 0 and 1, or in which only A's and not-A's are distinguished. We use this result later in **19.36**.

13.33. Reverting now to the more general case, suppose we have n families whose members number k_1, k_2, . . . k_n.

The ith family contributes $k_i(k_i-1)$ pairs to the intraclass table, and hence the total number of pairs is $S\{k_i(k_i-1)\} = N$, say, the summation extending over the n families.

Let the variate values be

$$\begin{array}{cccc} x_{11} \ x_{12} & . & . & . & x_{1k_1} \\ x_{21} \ x_{22} & . & . & . & x_{2k_2} \\ . & . & . \\ x_{n1} \ x_{n2} & . & . & . & x_{nk_n} \end{array}$$

As in **13.30**, we see that in the intraclass table each member of the first family appears (k_1-1) times, each of the second (k_2-1) times, and so on.

Hence,

$$\bar{X} = \bar{Y} = \frac{1}{N} S\{(k_i-1)S'(x_{ij})\} \quad . \qquad . \qquad . \quad (13.21)$$

the summation S' being carried over all members of the ith family and S over all families.

Similarly,

$$\sigma_x^2 = \sigma_Y^2 = \frac{1}{N} S''\{(k_i-1)S'(x_{ij}-\bar{X})^2\} \quad . \qquad . \quad (13.22)$$

and

$$\sigma^2 r = \frac{1}{N} S''\{(x_{ij}-\bar{X})(x_{im}-\bar{X})\}$$

the summation extending over all possible pairs.

17

and this, as in **13.31**, reduces to

$$N\sigma^2 r = S\{k_i^2 (\bar{X}_i - \bar{X})^2\} - SS'(x_{ij} - \bar{X})^2 \quad . \quad . \quad (13.23)$$

These formulæ are considerably more complex than (13.14), (13.16) and (13.20), but reduce to those forms if k_i is constant for all families.

SUMMARY.

1. In cases where the data are incomplete, or in order to avoid lengthy calculation, it is possible to use various methods of approximating to the product-moment coefficient of correlation, provided that the regression is approximately linear.

2. Cases in which the regression is non-linear can sometimes be reduced to the linear case by a suitable transformation of the variates.

3. The correlation ratio of X on Y is given by

$$\eta_{xy}^2 = 1 - \frac{\sigma_{ax}^2}{\sigma_x^2}$$

$$= \frac{\sigma_{mx}^2}{\sigma_x^2}$$

where σ_x^2 is the variance of X, σ_{ax}^2 is the weighted average of the variances of arrays and σ_{mx}^2 the variance of the means of X-arrays, weighted according to the number of individuals in the arrays.

4. $\eta_{xy}^2 - r^2$ cannot be negative, and if it is zero the regression of X on Y is linear.

5. The rank correlation coefficient is given by

$$\rho = \frac{S(xy)}{\sqrt{S(x^2)S(y^2)}}$$

where x and y are the deviations of the ranks X and Y from the mean $\frac{n+1}{2}$.

6. If

$$d_k = (X_k - Y_k)$$

$$\rho = 1 - \frac{6S(d^2)}{n^3 - n}$$

7. The coefficient of intraclass correlation is given by

$$\{1 + r(k - 1)\}\sigma^2 = k\sigma_m^2$$

where σ is the standard deviation of X and Y, and σ_m is the standard deviation of the means of families, there being n families each of k members.

EXERCISES.

13.1. Find to 3 places of decimals the correlation ratio of X on Y and of Y on X for the distribution of cows of Table 11.4, page 200 ($r = +0.219$). Hence, show that

$$\eta^2_{xy} - r^2 = 0.011$$
$$\eta^2_{yx} - r^2 = 0.023$$

13.2. Find the correlation ratios of the distribution of marriages of Table 11.2.

13.3. In a test of ability to distinguish shades of colour, 15 discs of various shades, whose true orders are 1, 2, . . . 15, are arranged by a subject in the order 7, 4, 2, 3, 1, 10, 6, 8, 9, 5, 11, 15, 14, 12, 13. Find the rank correlation coefficient between the real and the observed ranks.

13.4. Ten competitors in a beauty contest are ranked by three judges in the orders

1, 6, 5, 10, 3, 2, 4, 9, 7, 8
3, 5, 8, 4, 7, 10, 2, 1, 6, 9
6, 4, 9, 8, 1, 2, 3, 10, 5, 7

Use the rank correlation coefficient to discuss which pair of judges have the nearest approach to common tastes in beauty.

13.5. (*Cf*. Pearson, "On a Generalised Theory of Alternative Inheritance," *Phil. Trans.*, vol. 203, A, 1904, p. 53.) If we consider the correlation between number of recessive couplets in parent and in offspring, in a Mendelian population breeding at random (such as would ultimately result from an initial cross between a pure dominant and a pure recessive), the correlation is found to be $1/3$ for a total number of couplets n. If $n = 1$, the only possible numbers of recessive couplets are 0 and 1, and the correlation table between parent and offspring reduces to the form

Offspring.	Parent.		
	0	1	Total
0	5	1	6
1	1	1	2
Total	6	2	8

Verify the correlation, and work out the association coefficient Q.

13.6. (*Cf*. the above, and also Snow, *Proc. Roy. Soc.*, vol. 83, B, 1910, Table 3, p. 42.) For a similar population the correlation between brothers, assuming a practically infinite size of family, is $5/12$. The table is

Second Brother.	First Brother.		
	0	1	Total.
0	41	7	48
1	7	9	16
Total	48	16	64

Verify the correlation, and work out the association coefficient Q.

13.7. Referring to the notation of 13.25, show that we have the following expressions for the regressions in a fourfold table:—

$$r\frac{\sigma_1}{\sigma_2} = \frac{N\delta}{(B)(\beta)} = \frac{(AB)}{(B)} - \frac{(A\beta)}{(\beta)}$$

$$r\frac{\sigma_2}{\sigma_1} = \frac{N\delta}{(A)(a)} = \frac{(AB)}{(A)} - \frac{(aB)}{(a)}$$

Verify on the tables of Exercises 13.5 and 13.6.

13.8. In four pea-pods, each containing eight peas, the weights of the peas are, in hundredths of a gramme: 43, 46, 48, 42, 50, 45, 45 and 49; 33, 34, 37, 39, 32, 35, 37 and 41; 56, 52, 50, 51, 54, 52, 49 and 52; 36, 37, 38, 40, 40, 41, 44 and 44. Find the coefficient of intraclass correlation.

13.9. (Data from O. H. Latter, *Biometrika*, vol. 4, 1905, p. 363.)

The following table shows the length of cuckoos' eggs fostered by various birds:—

Length of Egg (units $\frac{1}{2}$ millimetre).

Foster Parent	40	41	42	43	44	45	46	47	48	49	50	Totals.
Robin	1	1	8	3	9	13	20	6	11	2	2	76
Wren	7	5	14	8	9	6	3	2	—	—	—	54
Hedge-Sparrow	—	—	2	5	14	13	13	3	5	—	3	58
Totals	8	6	24	16	32	32	36	11	16	2	5	188

Find the coefficient of intraclass correlation, and state how many entries there would be in the intraclass correlation table.

PARTIAL CORRELATION.

Multiple Correlation.

14.1. In Chapters 11 to 13 we developed the theory of the correlation between a single pair of variables. But in the case of statistics of attributes we found it necessary to proceed from the theory of simple association for a single pair of attributes to the theory of association for several attributes, in order to be able to deal with the complex causation characteristic of statistics ; and similarly the student will find it impossible to advance very far in the discussion of many problems in correlation without some knowledge of the theory of *multiple correlation*, or correlation between several variables.

For example, in considering the relationship between pauperism, out-relief and the age of recipients of relief, it might be found that changes in pauperism were highly correlated (positively) with changes in the out-relief ratio, and also with changes in the proportion of the old ; and the question might arise how far the first correlation was due merely to a tendency to give out-relief more freely to the old than the young, *i.e.* to a correlation between changes in out-relief and changes in proportion of the old. The question could not at the present stage be answered by working out the correlation coefficient between the last pair of variables, for we have as yet no guide as to how far a correlation between the variables 1 and 2 can be accounted for by correlations between 1 and 3 and 2 and 3.

Again, a marked positive correlation might be observed between, say, the bulk of a crop and the rainfall during a certain period, and practically no correlation between the crop and the accumulated temperature during the same period ; and the question might arise whether the last result might not be due merely to a negative correlation between rain and accumulated temperature, the crop being favourably affected by an increase of accumulated temperature *if other things were equal*, but failing as a rule to obtain this benefit owing to the concomitant deficiency of rain. In the problem of inheritance in a population, the corresponding problem is of great importance, as already indicated in Chapter 4. It is essential for the discussion of possible hypotheses to know whether an observed correlation between, say, grandson and grandparent can or cannot be accounted for solely by observed correlations between grandson and parent, parent and grandparent.

Partial Regressions and Correlation Coefficients.

14.2. Problems of this type, in which it is necessary to consider simultaneously the relations between at least three variables, and possibly

more, may be treated by a simple and natural extension of the method used in the case of two variables. The latter case was discussed by forming linear equations between the two variables, assigning such values to the constants as to make the sum of the squares of the errors of estimate as low as possible : the more complicated case may be discussed by forming linear equations between any one of the n variables involved, taking each in turn, and the $n-1$ others, again assigning such values to the constants as to make the sum of the squares of the errors of estimate a minimum. If the variables are X_1, X_2, X_3, . . . X_n, the equation will be of the form

$$X_1 = a + b_2 X_2 + b_3 X_3 + \ . \ . \ . \ + b_n X_n$$

If in such a **generalised regression equation** we find a sensible positive value for any one coefficient such as b_2, we know that there must be a positive correlation between X_1 and X_2 that cannot be accounted for by mere correlations of X_1 and X_2 with X_3, X_4 or X_n, for the effects of changes in these variables are allowed for in the remaining terms on the right. The magnitude of b_2 gives, in fact, the mean change in X_1 associated with a unit change in X_2 when all the remaining variables are kept constant.

The correlation between X_1 and X_2 indicated by b_2 may be termed a **partial correlation**, as corresponding with the *partial association* of Chapter 4, and it is required to deduce from the values of the coefficients b, which may be termed **partial regressions, partial coefficients of correlation** giving the correlation between X_1 and X_2 or other pair of variables *when the remaining variables X_3 . . . X_n are kept constant*, or when changes in these variables are corrected or allowed for, so far as this may be done with a linear equation. For examples of such generalised regression equations the student may turn to the illustrations worked out later (pp. 270–275).

14.3. With this explanatory introduction, we may now proceed to the algebraic theory of such generalised regression equations and of multiple correlation in general. It will first, however, be as well to revert briefly to the case of two variables. In Chapter 11, to obtain the greatest possible simplicity of treatment, the value of the coefficient $r = p/\sigma_1\sigma_2$ was deduced on the special assumption that the means of all arrays were strictly collinear, and the meaning of the coefficient in the more general case was subsequently investigated. Such a process is not conveniently applicable when a number of variables are to be taken into account, and the problem has to be faced directly : i.e. *required, to determine the coefficients and constant term, if any, in a regression equation, so as to make the sum of the squares of the errors of estimate a minimum.*

14.4. To solve this problem we proceed as in 11.20.

Let us measure the variates X_1 . . . X_n from their respective means, denoting the quantities so obtained by x_1 . . . x_n.

Then the regression equation of, say, x_1 on x_2 . . . x_n may be written in the form

$$x_1 = a_1 + b_2 x_2 + b_3 x_3 + \ . \ . \ . \ + b_n x_n$$

We have to find a_1, b_2, . . . b_n such that

$$E_1 = S(x_1 - a_1 - b_2 x_2 - \ . \ . \ . \ - b_n x_n)^2$$

is a minimum, the summation taking place over all sets of values of $x_1 \ldots x_n$.

Now,

$$E_1 = S(a_1{}^2) + S(x_1 - b_2 x_2 - \ldots - b_n x_n)^2$$

the product term

$$2S\{a_1(x_1 - b_2 x_2 - \ldots - b_n x_n)\}$$

vanishing, since x_1, etc. are measured from the mean.

Hence we have, for the minimum value of E_1,

$$a_1 = 0$$

Now, if b_2 is chosen so that E_1 is a minimum, the value of E_1, when $(b_2 + \delta)$ is substituted for b_2, is increased no matter how small δ may be ; *i.e.*

$$S\{x_1 - (b_2 + \delta)x_2 - \ldots - b_n x_n\}^2 \geqslant S(x_1 - b_2 x_2 - \ldots - b_n x_n)^2$$

Expanding the left-hand side, and neglecting δ^2, which can be made as small as we please compared with δ,

$$S(x_1 - b_2 x_2 - \ldots - b_n x_n)^2 - 2S\{x_2(x_1 - b_2 x_2 - \ldots - b_n x_n)\}\delta \geqslant S(x_1 - b_2 x_2 - \ldots - b_n x_n)^2$$

or

$$S\{x_2(x_1 - b_2 x_2 - \ldots - b_n x_n)\}\, \delta \leqslant 0$$

Now this is to be true for all small values of δ, positive or negative. If $S\{x_2(x_1 - b_2 x_2 - \ldots - b_n x_n)\}$ were not zero, this would be impossible, for if it were positive, say, we could take δ positive and the inequality would not be satisfied.

Hence,

$$S\{x_2(x_1 - b_2 x_2 - \ldots - b_n x_n)\} = 0$$

Similarly, considering b_3 instead of b_2, we have

$$S\{x_3(x_1 - b_2 x_2 - \ldots - b_n x_n)\} = 0$$

and so on, there being $(n - 1)$ equations. These are sufficient to determine the $(n - 1)$ quantities $b_2 \ldots b_n$, and hence our problem is solved.

Notation.

14.5. At this point we introduce a flexible notation which will enable us to consider any regression equation.

We write:

$$x_1 = b_{12.34 \ldots n} x_2 + b_{13.24 \ldots n} x_3 + \ldots + b_{1n.23 \ldots (n-1)} x_n \quad (14.1)$$

The quantities b are **partial regression coefficients**. The first subscript attached to the b is the subscript of the letter on the left (the dependent variable). The second subscript is that of the x to which it is attached. These are called **primary subscripts**.

After the primary subscripts, and separated from them by a point, are placed the subscripts of the remaining variables on the right. These are called **secondary subscripts**.

Equation (14.1) is the regression equation of x_1. Similarly, in accordance with the rules we have just laid down, we have :

$$x_2 = b_{21.34} \ldots {}_nx_1 + b_{23.14} \ldots {}_nx_3 + \ldots + b_{2n.13} \ldots {}_{(n-1)}x_n$$

and so on.

It should be noted that the order in which the secondary subscripts are written is immaterial; but this is not true of the primary subscripts; *e.g.* $b_{12.3} \ldots {}_n$ and $b_{21.3} \ldots {}_n$ denote quite distinct coefficients, x_1 being the dependent variable in the first case and x_2 in the second.

A coefficient with p secondary subscripts may be termed a **regression of the pth order.** The regressions b_{12}, b_{21}, b_{13}, b_{31}, etc., obtained by considering two variables alone, may be regarded as of order zero, and may be termed total, as distinct from **partial,** regressions.

14.6. If the regressions $b_{12.34} \ldots {}_n$, $b_{13.24} \ldots {}_n$, etc., be assigned the " best " values, as determined by the method of least squares, the difference between the actual value of x_1 and the value assigned by the right-hand side of the regression equation (14.1), that is, the error of estimate, will be denoted by $x_{1.23} \ldots {}_n$; *i.e.* as a definition we have

$$x_{1.23} \ldots {}_n = x_1 - b_{12.34} \ldots {}_nx_2 - b_{13.24} \ldots {}_nx_3 - \ldots - b_{1n.23} \ldots {}_{(n-1)}x_n \quad (14.2)$$

where x_1, x_2, $\ldots$ x_n are assigned any one set of observed values. Such an error (or **residual,** as it is sometimes called), denoted by a symbol with p secondary suffixes, will be termed a **deviation** of the pth order.

Finally, we will define a generalised standard deviation $\sigma_{1.23} \ldots {}_n$ by the equation

$$N\sigma_{1.23}^2 \ldots {}_n = S(x_{1.23}^2 \ldots {}_n) \quad . \quad . \quad . \quad (14.3)$$

N being, as usual, the number of observations. A standard deviation denoted by a symbol with p secondary suffixes will be termed a standard deviation of the pth order, the standard deviations σ_1, σ_2, etc., being regarded as of order zero, the standard deviations $\sigma_{1.2}$, $\sigma_{2.1}$, etc., of the first order, and so on.

14.7. In the case of two variables, the correlation coefficient r_{12} may be regarded as defined by the equation

$$r_{12} = (b_{12}b_{21})^{\frac{1}{2}}$$

We shall generalise this equation in the form

$$r_{12.34} \ldots {}_n = (b_{12.34} \ldots {}_nb_{21.34} \ldots {}_n)^{\frac{1}{2}} \quad . \quad . \quad (14.4)$$

This is at present a pure definition of a new symbol, and it remains to be shown that $r_{12.34} \ldots {}_n$ may really be regarded as, and possesses all the properties of, a correlation coefficient; the name may, however, be applied to it, pending the proof. A correlation coefficient with p secondary subscripts will be termed a **correlation of order p.** Evidently, in the case of a correlation coefficient, the order in which both primary and secondary subscripts is written is indifferent, for the right-hand side of equation (14.4) is unaltered by writing 2 for 1 and 1 for 2. The correlations r_{12}, r_{13}, etc., may be regarded as of order zero, and spoken of as total, as distinct from **partial,** correlations.

The Normal Equations.

14.8. All the quantities we have just defined are expressible in terms of the total and partial regression coefficients, and particular importance therefore attaches to the equations which give those coefficients. The equations of **14.4** may be written

$$S(x_2 x_{1.23 \ldots n}) = 0 \qquad . \qquad . \qquad . \qquad (14.5)$$

etc., there being $(n-1)$ equations for each regression equation.

These equations are called the **normal equations.** We shall see below that in practical cases it is usually more convenient not to solve them directly but to proceed in stages, finding first the regressions and correlations of order zero, then those of order 1, and so on.

14.9. If the student will follow the process by which (14.5) was obtained, he will see that when the condition is expressed that $b_{12.34 \ldots n}$ shall possess the "least-square" value, x_2 enters into the product-sum with $x_{1.23 \ldots n}$; when the same condition is expressed for $b_{13.24 \ldots n}$, x_3 enters into the product-sum, and so on. Taking each regression in turn, in fact, every x the suffix of which is included in the secondary suffixes of $x_{1.23 \ldots n}$ enters into the product-sum. The normal equations of the form (14.5) are therefore equivalent to the theorem :

The product-sum of any deviation of order zero with any deviation of higher order is zero, provided the subscript of the former occur among the secondary subscripts of the latter.

14.10. But it follows from this that

$$S(x_{1.34 \ldots n} x_{2.34 \ldots n}) = S\{x_{1.34 \ldots n}(x_2 - b_{23.4 \ldots n}x_3 - \ldots - b_{2n.34 \ldots (n-1)}x_n)\}$$
$$= S(x_{1.34 \ldots n} x_2)$$

Similarly,

$$S(x_{1.34 \ldots n} x_{2.34 \ldots n}) = S(x_1 x_{2.34 \ldots n})$$

Similarly again,

$$S(x_{1.34 \ldots n} x_{2.34 \ldots (n-1)}) = S(x_{1.34 \ldots n} x_2)$$

and so on. Therefore, quite generally,

$$\left.
\begin{aligned}
S(x_{1.34 \ldots n} x_{2.34 \ldots n}) &= S(x_{1.34 \ldots (n-1)} x_{2.34 \ldots n}) \\
&= \quad . \qquad . \qquad . \qquad . \\
&= S(x_1 x_{2.34 \ldots n}) \\
&= S(x_{1.34 \ldots n} x_{2.34 \ldots (n-1)}) \\
&= \quad . \qquad . \qquad . \qquad . \\
&= S(x_{1.34 \ldots n} x_2)
\end{aligned}
\right\} \qquad . \qquad . \qquad . \qquad (14.6)$$

Comparing all the equal product-sums that may be obtained in this way, we see that *the product-sum of any two deviations is unaltered by omitting any or all of the secondary subscripts of either which are common to the two, and, conversely, the product-sum of any deviation of order p with a deviation of order p + q, the p subscripts being the same in each case, is unaltered by adding to the secondary subscripts of the former any or all of the q additional subscripts of the latter.*

It follows therefore from (14.5) that *any product-sum is zero if all the subscripts of the one deviation occur among the secondary subscripts of the other.* As the simplest case, we may note that x_1 is uncorrelated with $x_{2.1}$, and x_2 uncorrelated with $x_{1.2}$.

The theorems of this and of the preceding paragraph are of fundamental importance, and should be carefully remembered.

14.11. We can now show that the quantities r defined by (14.4) are really coefficients of correlation. In fact we have, from the results of **14.9** and **14.10**,

$$0 = S(x_{2.34 \ldots n}x_{1.234 \ldots n})$$
$$= S\{x_{2.34 \ldots n}(x_1 - b_{12.34 \ldots n}x_2 - \text{terms in } x_3 \text{ to } x_n)\}$$
$$= S(x_1 x_{2.34 \ldots n}) - b_{12.34 \ldots n}S(x_2 x_{2.34 \ldots n})$$
$$= S(x_{1.34 \ldots n}x_{2.34 \ldots n}) - b_{12.34 \ldots n}S(x^2_{2.34 \ldots n})$$

That is,

$$b_{12.34 \ldots n} = \frac{S(x_{1.34 \ldots n}x_{2.34 \ldots n})}{S(x^2_{2.34 \ldots n})} \qquad . \qquad . \quad (14.7)$$

But this is the value that would have been obtained by taking a regression equation of the form

$$x_{1.34 \ldots n} = b_{12.34 \ldots n}x_{2.34 \ldots n}$$

and determining $b_{12.34 \ldots n}$ by the method of least squares, *i.e.* $b_{12.34 \ldots n}$ is the regression of $x_{1.34 \ldots n}$ on $x_{2.34 \ldots n}$. It follows at once from (14.4) that $r_{12.34 \ldots n}$ is the correlation between $x_{1.34 \ldots n}$ and $x_{2.34 \ldots n}$, and from (14.3) that we may write

$$b_{12.34 \ldots n} = r_{12.34 \ldots n}\frac{\sigma_{1.34 \ldots n}}{\sigma_{2.34 \ldots n}} \qquad . \qquad . \quad (14.8)$$

an equation identical with the familiar relation $b_{12} = r_{12}\sigma_1/\sigma_2$, with the secondary suffixes $34 \ldots n$ added throughout.

To illustrate the meaning of the equation by the simplest case, if we had three variables only, x_1, x_2 and x_3, the value of $b_{12.3}$ or $r_{12.3}$ could be determined (1) by finding the correlations r_{13} and r_{23} and the corresponding regressions b_{13} and b_{23}; (2) working out the residuals $x_1 - b_{13}x_3$ and $x_2 - b_{23}x_3$ for all associated deviations; (3) working out the correlation between the residuals associated with the same values of x_3. The method would not, however, be a practical one, as the arithmetic would be extremely lengthy, much more lengthy than the method given below for expressing a correlation of order p in terms of correlations of order $p - 1$.

Expression of Standard Deviation in terms of Standard Deviations and Coefficients of Lower Orders.

14.12. *Any standard deviation of order* p *may be expressed in terms of a standard deviation of order* p $-$ 1 *and a correlation of order* p $-$ 1. For,

$$S(x_{1.23 \ldots n})^2 = S(x_{1.23 \ldots (n-1)}x_{1.23 \ldots n})$$
$$= S(x_{1.23 \ldots (n-1)})(x_1 - b_{1n.23 \ldots (n-1)}x_n - \text{terms in } x_2 \text{ to } x_{n-1})$$
$$= S(x^2_{1.23 \ldots (n-1)}) - b_{1n.23 \ldots (n-1)}S(x_{1.23 \ldots (n-1)}x_{n.23 \ldots (n-1)})$$

or, dividing through by the number of observations:

$$\sigma^2_{1.23 \ldots n} = \sigma^2_{1.23 \ldots (n-1)}(1 - b_{1n.23 \ldots (n-1)}b_{n1.23 \ldots (n-1)})$$

$$= \sigma^2_{1.23 \ldots (n-1)}(1 - r^2_{1n.23 \ldots (n-1)}) \quad . \quad . \quad . \quad (14.9)$$

This is again the relation of the familiar form

$$\sigma^2_{1.n} = \sigma^2_1(1 - r^2_{1n})$$

with the secondary suffixes $23 \ldots (n-1)$ added throughout. It is clear from (14.9) that $r_{1n.23 \ldots (n-1)}$, like any correlation of order zero, cannot be numerically greater than unity. It also follows at once that if we have been estimating x_1 from $x_2, x_3, \ldots x_{n-1}, x_n$ will not increase the accuracy of estimate unless $r_{1n.23 \ldots (n-1)}$ (not r_{1n}) differ from zero. This condition is somewhat interesting, as it leads to rather unexpected results. For example, if $r_{12} = +0.8$, $r_{13} = +0.4$, $r_{23} = +0.5$, it will not be possible to estimate x_1 with any greater accuracy from x_2 and x_3 than from x_2 alone, for the value of $r_{13.2}$ is zero (see below, **14.15**).

14.13. It should be noted that, in equation (14.9), any other subscript can be eliminated in the same way as subscript n from the suffix of $\sigma_{1.23 \ldots n}$, so that a standard deviation of order p can be expressed in p ways in terms of standard deviations of the next lower order. This is useful as affording an independent check on arithmetic. Further, $\sigma_{1.23 \ldots (n-1)}$ can be expressed in the same way in terms of $\sigma_{1.23 \ldots (n-2)}$, and so on, so that we must have

$$\sigma^2_{1.23 \ldots n} = \sigma^2_1(1 - r^2_{12})(1 - r^2_{13.2})(1 - r^2_{14.23}) \ldots (1 - r^2_{1n.23 \ldots (n-1)}) \quad (14.10)$$

This is an extremely convenient expression for arithmetical use; the arithmetic can again be subjected to an absolute check by eliminating the subscripts in a different, say the inverse, order. Apart from the algebraic proof, it is obvious that the values must be identical; for if we are estimating one variable from n others, it is clearly indifferent in what order the latter are taken into account.

$\sigma_{1.23 \ldots n}$ can also be expressed in terms of σ_1 and the total correlation coefficients. We have

$$S(x_{1.23 \ldots n})^2 = S\{x_1(x_{1.23 \ldots n})\} = N\sigma^2_{1.23 \ldots n}$$

Hence, expanding $x_{1.23 \ldots n}$,

$$\sigma^2_1 - b_{12.3 \ldots n}r_{12}\sigma_1\sigma_2 - b_{13.2 \ldots n}r_{13}\sigma_1\sigma_3 - \ldots = \sigma^2_{1.23 \ldots n}$$

The $(n-1)$ normal equations involving $x_{1.23 \ldots n}$ are

$$S(x_2 x_{1.23 \ldots n}) = 0, \text{ etc.}$$

i.e. expanding,

$$r_{21}\sigma_1\sigma_2 - b_{12.3 \ldots n}\sigma^2_2 - b_{13.2 \ldots n}r_{23}\sigma_2\sigma_3 \ldots = 0$$

$$r_{31}\sigma_1\sigma_3 - b_{12.3 \ldots n}r_{32}\sigma_3\sigma_2 - b_{13.2 \ldots n}\sigma^2_3 \ldots = 0, \text{ etc.}$$

Regarding the n equations so obtained as equations in the quantities b, we have, on elimination, the determinant

$$\begin{vmatrix} \sigma_1^2 - \sigma_{1.23}^2 \ldots n & r_{12}\sigma_1\sigma_2 & r_{13}\sigma_1\sigma_3 \cdots r_{1n}\sigma_1\sigma_n \\ r_{21}\sigma_2\sigma_1 & \sigma_2^2 & r_{23}\sigma_2\sigma_3 \cdots r_{2n}\sigma_2\sigma_n \\ \cdot\cdot & \cdot\cdot & \cdot\cdot \qquad \cdot\cdot \\ r_{n1}\sigma_n\sigma_1 & r_{n2}\sigma_n\sigma_2 & r_{n3}\sigma_n\sigma_3 \cdots \sigma_n^2 \end{vmatrix} = 0$$

Dividing the sth row by σ_s and the tth column by σ_t, this gives:

$$\begin{vmatrix} 1 - \dfrac{\sigma_{1.23}^2 \ldots n}{\sigma_1^2} & r_{12} & r_{13} \cdots r_{1n} \\ r_{21} & 1 & r_{23} \cdots r_{2n} \\ \cdot\cdot & \cdot\cdot & \cdot\cdot \qquad \cdot\cdot \\ r_{n1} & r_{n2} & r_{n3} \cdots 1 \end{vmatrix} = 0$$

Write ω for the determinant

$$\begin{vmatrix} 1 & r_{12} \cdots r_{1n} \\ r_{21} & 1 \cdots r_{2n} \\ \cdot & \cdot \quad \cdot \quad \cdot \\ r_{n1} & r_{n2} \cdots 1 \end{vmatrix}$$

and let ω_{11} be the minor of the term in the first row and column.   Then

$$\omega - \frac{\sigma_{1.23}^2 \ldots n}{\sigma_1^2}\,\omega_{11} = 0$$

$$\frac{\sigma_{1.23}^2 \ldots n}{\sigma_1^2} = \frac{\omega}{\omega_{11}} \qquad . \qquad . \qquad . \quad (14.11)$$

Similarly,

$$\frac{\sigma_{2.13}^2 \ldots n}{\sigma_2^2} = \frac{\omega}{\omega_{22}}$$

and so on.

These results exhibit $\sigma_{1.23}^2 \ldots {}_n$, etc., in a symmetrical form.

Expression of Regression Coefficients in terms of Coefficients of Lower Orders.

14.14.   Any regression of order p may be expressed in terms of regressions of order $p - 1$.   For we have:

$$S(x_{1.34\ldots n}x_{2.34\ldots n}) = S(x_{1.34\ldots(n-1)}x_{2.34\ldots n})$$
$$= S(x_{1.34\ldots(n-1)})(x_2 - b_{2n.34\ldots(n-1)}x_n - \text{terms in } x_3 \text{ to } x_{n-1})$$
$$= S(x_{1.34\ldots(n-1)}x_{2.34\ldots(n-1)}) - b_{2n.34\ldots(n-1)}S(x_{1.34\ldots(n-1)}x_{n.34\ldots(n-1)})$$

Replacing $b_{2n.34\ldots(n-1)}$ by $b_{n2.34\ldots(n-1)}\sigma_{2.34\ldots(n-1)}^2/\sigma_{n.34\ldots(n-1)}^2$

we have:

$$b_{12.34\ldots n}\sigma_{2.34\ldots n}^2 = b_{12.34\ldots(n-1)}\sigma_{2.34\ldots(n-1)}^2 - b_{1n.34\ldots(n-1)}b_{n2.34\ldots(n-1)}\sigma_{2.34\ldots(n-1)}^2$$

or, from (14.9),

$$b_{12.34 \ldots n} = \frac{b_{12.34 \ldots (n-1)} - b_{1n.34 \ldots (n-1)} b_{n2.34 \ldots (n-1)}}{1 - b_{2n.34 \ldots (n-1)} b_{n2.34 \ldots (n-1)}} \qquad (14.12)$$

The student should note that this is an expression of the form

$$b_{12.n} = \frac{b_{12} - b_{1n} b_{n2}}{1 - b_{2n} b_{n2}}$$

with the subscripts $34 \ldots (n-1)$ added throughout. The coefficient $b_{12.34 \ldots n}$ may therefore be regarded as determined from a regression equation of the form

$$x_{1.34 \ldots (n-1)} = b_{12.34 \ldots n} x_{2.34 \ldots (n-1)} + b_{1n.23 \ldots (n-1)} x_{n.34 \ldots (n-1)}$$

i.e. it is the partial regression of $x_{1.34 \ldots (n-1)}$ on $x_{2.34 \ldots (n-1)}$, $x_{n.34} \ldots (n-1)$ being given. As any other secondary suffix might have been eliminated in lieu of n, we might also regard it as the partial regression of $x_{1.45 \ldots n}$ on $x_{2.45 \ldots n}$, $x_{3.45 \ldots n}$ being given, and so on.

Expression of Correlation Coefficient in terms of Coefficients of Lower Orders.

14.15. From equation (14.12) we may readily obtain a corresponding equation for correlations. For (14.12) may be written:

$$b_{12.34 \ldots n} = \frac{r_{12.34 \ldots (n-1)} - r_{1n.34 \ldots (n-1)} r_{2n.34 \ldots (n-1)}}{1 - r_{2n.34 \ldots (n-1)}^2} \cdot \frac{\sigma_{1.34 \ldots (n-1)}}{\sigma_{2.34 \ldots (n-1)}}$$

Hence, writing down the corresponding expression for $b_{21.34 \ldots n}$ and taking the square root:

$$r_{12.34 \ldots n} = \frac{r_{12.34 \ldots (n-1)} - r_{1n.34 \ldots (n-1)} r_{2n.34 \ldots (n-1)}}{(1 - r_{1n.34 \ldots (n-1)}^2)^{\frac{1}{2}} (1 - r_{2n.34 \ldots (n-1)}^2)^{\frac{1}{2}}} \qquad (14.13)$$

This is, similarly, the expression for three variables:

$$r_{12.n} = \frac{r_{12} - r_{1n} r_{2n}}{(1 - r_{1n}^2)^{\frac{1}{2}} (1 - r_{2n}^2)^{\frac{1}{2}}}$$

with the secondary subscripts added throughout, and $r_{12.34 \ldots n}$ can be assigned interpretations corresponding to those of $b_{12.34 \ldots n}$ above. Evidently equation (14.13) permits of an absolute check on the arithmetic in the calculation of all partial coefficients of an order higher than the first, for any one of the secondary suffixes of $r_{12.34 \ldots n}$ can be eliminated so as to obtain another equation of the same form as (14.13), and the value obtained for $r_{12.34 \ldots n}$ by inserting the values of the coefficients of lower order in the expression on the right must be the same in each case.

Practical Procedure.

14.16. The equations now obtained provide all that is necessary for the arithmetical solution of problems in multiple correlation. The best mode of procedure on the whole, having calculated all the correlations and standard deviations of order zero, is (1) to calculate the correlations

of higher order by successive applications of equation (14.13); (2) to calculate any required standard deviations by equation (14.10); (3) to calculate any required regressions by equation (14.8); the use of equation (14.12) for calculating the regressions of successive orders directly from one another is comparatively clumsy. We will give two illustrations, the first for three and the second for four variables. The introduction of more variables does not involve any difference in the form of the arithmetic, but rapidly increases the amount.[1]

Example 14.1.—In Exercise 11.2, page 224, we gave some data of (1) the average earnings of agricultural labourers, (2) the percentage of the population in receipt of poor law relief, (3) the ratios of the numbers in receipt of outdoor relief to those relieved in the workhouse, for 38 rural districts. ·Required to work out the partial correlations, regressions, etc., for these three variables.

Using as our notation X_1 = average earnings, X_2 = percentage of population in receipt of relief, X_3 = out-relief ratio, the first constants determined are:

$$M_1 = 15 \cdot 9 \text{ shillings} \qquad \sigma_1 = 1 \cdot 71 \text{ shillings} \qquad r_{12} = -0 \cdot 66$$
$$M_2 = 3 \cdot 67 \text{ per cent.} \qquad \sigma_2 = 1 \cdot 29 \text{ per cent.} \qquad r_{13} = -0 \cdot 13$$
$$M_3 = 5 \cdot 79 \qquad \sigma_3 = 3 \cdot 09 \qquad r_{23} = +0 \cdot 60$$

To obtain the partial correlations, equation (14.13) is used direct in its simplest form:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{(1 - r_{13}^2)^{\frac{1}{2}}(1 - r_{23}^2)^{\frac{1}{2}}}$$

The work is best done systematically and the results collected in tabular form, especially if logarithms are used, as many of the logarithms occur repeatedly. First, it will be noted that the logarithms of $(1 - r^2)^{\frac{1}{2}}$ occur in all the denominators; these had, accordingly, better be worked out at once and tabulated (col. 2 of the table below). In column 3 the

1.	2.	3.	4.	5.	6.	7.	8.		9.
						Correlation of First Order.			
	$\log \sqrt{1-r^2}$	Product Term.	Numerator.	log Num.	log Denom.	log.	Value.		$\log \sqrt{1-r^2}$
$r_{12} = -0 \cdot 66$	$\bar{1} \cdot 87580$	$-0 \cdot 0780$	$-0 \cdot 5820$	$\bar{1} \cdot 76492$	$\bar{1} \cdot 89938$	$\bar{1} \cdot 86554$	$r_{12 \cdot 3} -0 \cdot 73$		$\bar{1} \cdot 83216$
$r_{13} = -0 \cdot 13$	$\bar{1} \cdot 99629$	$-0 \cdot 3960$	$+0 \cdot 2660$	$\bar{1} \cdot 42485$	$\bar{1} \cdot 77889$	$\bar{1} \cdot 64599$	$r_{13 \cdot 2} +0 \cdot 44$		$\bar{1} \cdot 95267$
$r_{23} = +0 \cdot 60$	$\bar{1} \cdot 90309$	$+0 \cdot 0858$	$+0 \cdot 5142$	$\bar{1} \cdot 71113$	$\bar{1} \cdot 87209$	$\bar{1} \cdot 83904$	$r_{23 \cdot 1} +0 \cdot 69$		$\bar{1} \cdot 85946$

product term of the numerator of each partial coefficient is entered, *i.e.* the product of the two other coefficients on the remaining lines in column 1; subtracting this from the coefficient on the same line in column 1, we have the numerator (col. 4) and can enter its logarithm. The logarithm of the

---

[1] It will he noticed from the preceding work that all correlations are assumed to be determined by the product-sum formula. The method has been applied with correlations determined in other ways, *e.g.* from fourfold or contingency tables or by the method of ranks. In spite of the favourable result of an experimental test (Ethel M. Newbold, "Notes on an Experimental Test of Errors in Partial Correlation derived from Four-fold and Biserial Total Coefficients," *Biometrika*, vol. 17, 1925, p. 251), the results obtained in such ways remain of doubtful value.

denominator (col. 6) is obtained at once by adding the two logarithms of $(1-r^2)^{\frac{1}{2}}$ on the remaining lines of the table, and subtracting the logarithms of the denominators from those of the numerators, we have the logarithms of the correlations of the first order. It is also as well to calculate at once, for reference in the calculation of standard deviations of the second order, the values of $\log \sqrt{1-r^2}$ for the first-order coefficients (col. 9).

Having obtained the correlations, we can now proceed to the regressions. If we wish to find all the regression equations, we shall have six regressions to calculate from equations of the form

$$b_{12.3} = r_{12.3}\sigma_{1.3}/\sigma_{2.3}$$

These will involve all the six standard deviations of the first order $\sigma_{1.2}$, $\sigma_{1.3}$, $\sigma_{2.1}$, $\sigma_{2.3}$, etc. The standard deviations of the first order are not in themselves of much interest, but the standard deviations of the second order are important, as being the standard errors or root-mean-square errors of estimate made in using the regression equations of the second order. We may save needless arithmetic, therefore, by replacing the standard deviations of the first order by those of the second, omitting the former entirely, and transforming the above equation for $b_{12.3}$ to the form

$$b_{12.3} = r_{12.3}\sigma_{1.23}/\sigma_{2.13}$$

This transformation is a useful one and should be noted by the student. The values of each σ may be calculated twice independently by the formulæ of the form

$$\sigma_{1.23} = \sigma_1(1-r_{12}^2)^{\frac{1}{2}}(1-r_{13.2}^2)^{\frac{1}{2}}$$
$$= \sigma_1(1-r_{13}^2)^{\frac{1}{2}}(1-r_{12.3}^2)^{\frac{1}{2}}$$

so as to check the arithmetic; the work is rapidly done if the values of $\log \sqrt{1-r^2}$ have been tabulated. The values found are :

$$\log \sigma_{1.23} = 0\cdot06146 \qquad \sigma_{1.23} = 1\cdot15$$
$$\log \sigma_{2.13} = \bar{1}\cdot84584 \qquad \sigma_{2.13} = 0\cdot70$$
$$\log \sigma_{3.12} = 0\cdot34571 \qquad \sigma_{3.12} = 2\cdot22$$

From these and the logarithms of the r's we have :

$\log b_{12.3} = 0\cdot08116,$	$b_{12.3} = -1\cdot21$	$\log b_{13.2} = \bar{1}\cdot36174,$	$b_{13.2} = +0\cdot23$
$\log b_{21.3} = \bar{1}\cdot64993,$	$b_{21.3} = -0\cdot45$	$\log b_{23.1} = \bar{1}\cdot33917,$	$b_{23.1} = +0\cdot22$
$\log b_{31.2} = \bar{1}\cdot93024,$	$b_{31.2} = +0\cdot85$	$\log b_{32.1} = 0\cdot33891,$	$b_{32.1} = +2\cdot18$

That is, the regression equations are :

(1) $x_1 = -1\cdot21x_2 + 0\cdot23x_3$
(2) $x_2 = -0\cdot45x_1 + 0\cdot22x_3$
(3) $x_3 = +0\cdot85x_1 + 2\cdot18x_2$

or, transferring the origins to zero : .

(1) *Earnings* $\qquad X_1 = +19\cdot0 - 1\cdot21X_2 + 0\cdot23X_3$
(2) *Pauperism* $\qquad X_2 = +9\cdot55 - 0\cdot45X_1 + 0\cdot22X_3$
(3) *Out-relief ratio* $X_3 = -15\cdot7 + 0\cdot85X_1 + 2\cdot18X_2$

The units are throughout one shilling for the earnings X_1, 1 per cent. for the pauperism X_2 and 1 for the out-relief ratio X_3.

Now let us examine the light thrown by these results on the relationship between the variables.

The first and second regression equations are those of most practical importance. The argument has been advanced that the giving of out-relief tends to lower earnings, and the total coefficient $(r_{13} = -0.13)$ between earnings (X_1) and out-relief (X_3), though very small, does not seem inconsistent with such a hypothesis. The partial correlation coefficient $(r_{13.2} = +0.44)$ and the regression equation (1), however, indicate that in unions with a *given* percentage of the population in receipt of relief (X_2) the earnings are highest where the proportion of out-relief is highest; and this is, in so far, against the hypothesis of a tendency to lower wages. It remains possible, of course, that out-relief may adversely affect the *possi-bility of earning*, e.g. by limiting the employment of the old.

As regards pauperism, the argument might be advanced that the observed correlation $(r_{23} = +0.60)$ between pauperism and out-relief was in part due to the negative correlation $(r_{13} = -0.13)$ between earnings and out-relief. Such a hypothesis would have little to support it in view of the smallness and doubtful significance of r_{13}, and is definitely contradicted by the positive partial correlation $r_{23.1} = +0.69$ and the second regression equation. The third regression equation shows that the proportion of out-relief is on the whole highest where earnings are highest and pauperism greatest. It should be noticed, however, that a negative ratio is clearly impossible, and consequently the relation cannot be strictly linear; but the third equation gives *possible* (positive) average ratios for all the combinations of pauperism and earnings that actually occur.

Example 14.2 (Four Variables).—As an illustration of the form of the work in the case of four variables, we will take a portion of the data from another investigation into the causation of pauperism.

The variables are the ratios of the values in 1891 to the values in 1881 (taken as 100) of—

  1. The percentage of the population in receipt of relief,
  2. The ratio of the numbers given outdoor relief to the numbers relieved in the workhouse,
  3. The percentage of the population over 65 years of age,
  4. The population itself,

in the metropolitan group of 32 unions, and the fundamental constants (means, standard deviations and correlations) are as follows :—

TABLE 14.1.

1. Means.		2. Standard deviations.		3. Correlation coefficient.		4. $\log \sqrt{1 - r^2}$.
1	104·7	1	29·2	12	+0·52	$\bar{1}$·93154
2	90·6	2	41·7	13	+0·41	$\bar{1}$·96003
3	107·7	3	5·5	14	−0·14	$\bar{1}$·99570
4	111·3	4	23·8	23	+0·49	$\bar{1}$·94038
—	—	—	—	24	+0·23	$\bar{1}$·98820
—	—	—	—	34	+0·25	$\bar{1}$·98598

It is seen that the average changes are not great ; the percentages of the population in receipt of relief have increased on an average by 4·7 per cent., the out-relief ratio has dropped by 9·4 per cent. and the percentage of the old has increased by 7·7 per cent., while the population of the unions has risen on the average by 11·3 per cent. At the same time the standard deviations of the first, second and fourth variables are very large. As a matter of fact, while in one union the pauperism decreased by nearly 50 per cent. and in others by 20 per cent., in some there were increases of

TABLE 14.2.

1.		2.	3.	4.		5.
Correlation coefficient (Zero Order).		Product Term of Numerator.	Numerator.	Correlation coefficient (First Order).		$\log \sqrt{1-r^2}$.
12	+0·52	+0·2009	+0·3191	12·3	+0·4013	$\bar{1}$·96187
13	+0·41	+0·2548	+0·1552	13·2	+0·2084	$\bar{1}$·99035
23	+0·49	+0·2132	·+0·2768	23·1	+0·3553	$\bar{1}$·97070
12	+0·52	−0·0322	+0·5522	12·4	+0·5731	$\bar{1}$·91355
14	−0·14	+0·1196	−0·2596	14·2	−0·3123	$\bar{1}$·97772
24	+0·23	−0·0728	+0·3028	24·1	+0·3580	$\bar{1}$·97022
13	+0·41	−0·0350	+0·4450	13·4	+0·4642	$\bar{1}$·94731
14	−0·14	+0·1025	−0·2425	14·3	.−0·2746	$\bar{1}$·98297
34	+0·25	−0·0574	+0·3074	34·1	+0·3404	$\bar{1}$·97326
23	+0·49	+0·0575	+0·4325	23·4	+0·4590	$\bar{1}$·94863
24	+0·23	+0·1225	+0·1075	24·3	+0·1274	$\bar{1}$·99645
34	+0·25	+0·1127	+0·1373	34·2	+0·1618	1·99424

60, 80 and 90 per cent. ; similarly, in the case of the out-relief, in several unions the ratio was decreased by 40 to 60 per cent., a consistent anti-out-relief policy having been enforced ; in others the ratio was doubled, and more than doubled. As regards population, the more central districts showed decreases ranging up to 20 and 25 per cent., the circumferential districts increases of 45 to 80 per cent. The correlations of order zero are not large, the changes in the rate of pauperism exhibiting the highest correlation with changes in the out-relief ratio, slightly less with changes in the proportion of old and very little with changes in population.

The correlations of the second order are obtained in two steps. In the first place, the six coefficients of order zero are grouped in four sets of three, corresponding to the four sets of three variables formed by omitting each one of the four variables in turn (Table 14.2, col. 1). Each of these sets of three coefficients is then treated in the same manner as in the last example, and so the correlations of the first order (Table 14.2, col. 4) are obtained. The first-order coefficients are then regrouped in sets of three, with the same secondary suffix (Table 14.3, col. 1), and these are treated precisely in the same way as the coefficients of order zero. In this way, it will be seen, the value of each coefficient of the second order is arrived at in two ways independently, and so the arithmetic is checked : $r_{12.34}$ occurs in

18

the first and fourth lines, for instance, $r_{13.24}$ in the second and seventh, and so on. Of course slight differences may occur in the last digit if a sufficient number of digits is not retained, and for this reason the intermediate work should be carried to a greater degree of accuracy than is necessary in the final result ; thus four places of decimals were retained throughout in the intermediate work of this example, and three in the final result. If he carries out an independent calculation, the student may differ slightly from the logarithms given in this and the following work, if more or fewer figures are retained.

TABLE 14.3.

1. Correlation coefficient (First Order).		2. Product Term of Numerator.	3. Numerator.	4. Correlation coefficient (Second Order).		5. $\log \sqrt{1-r^2}$.
12·4	+0·5731	+0·2131	+0·3600	12·34	+0·457	$\bar{1}$·94901
13·4	+0·4642	+0·2631	+0·2011	13·24	+0·276	$\bar{1}$·98277
23·4	+0·4590	+0·2660	+0·1930	23·14	+0·266	$\bar{1}$·98408
12·3	+0·4013	−0·0350	+0·4363	12·34	+0·457	—
14·3	−0·2746	+0·0511	−0·3257	14·23	−0·359	$\bar{1}$·97013
24·3	+0·1274	−0·1102	+0·2376	24·13	+0·270	$\bar{1}$·98359
13·2	+0·2084	−0·0505	+0·2589	13·24	+0·276	—
14·2	−0·3123	+0·0337	−0·3460	14·23	−0·359	—
34·2	+0·1618	−0·0651	+0·2269	34·12	+0·244	$\bar{1}$·98664
23·1	+0·3553	+0·1219	+0·2334	23·14	+0·266	—
24·1	+0·3580	+0·1209	+0·2371	24·13	+0·270	—
34·1	+0·3404	+0·1272	+0·2132	34·12	+0·244	—

Having obtained the correlations, the regressions can be calculated from the third-order standard deviations by equations of the form (as in the last example),

$$b_{12.34} = r_{12.34} \frac{\sigma_{1.234}}{\sigma_{2.134}}$$

so the standard deviations of lower orders need not be evaluated. Using equations of the form

$$\sigma_{1.234} = \sigma_1 (1 - r_{12}^2)^{\frac{1}{2}} (1 - r_{13.2}^2)^{\frac{1}{2}} (1 - r_{14.23}^2)^{\frac{1}{2}}$$
$$= \sigma_1 (1 - r_{14}^2)^{\frac{1}{2}} (1 - r_{13.4}^2)^{\frac{1}{2}} (1 - r_{12.34}^2)^{\frac{1}{2}}$$

we find :

$$\log \sigma_{1.234} = 1·35740 \qquad \sigma_{1.234} = 22·8$$
$$\log \sigma_{2.134} = 1·50597 \qquad \sigma_{2.134} = 32·1$$
$$\log \sigma_{3.124} = 0·65773 \qquad \sigma_{3.124} = 4·55$$
$$\log \sigma_{4.123} = 1·32914 \qquad \sigma_{4.123} = 21·3$$

All the twelve regressions of the second order can be readily calculated, given these standard deviations and the correlations, but we may confine

ourselves to the equation giving the changes in pauperism (X_1) in terms of other variables as the most important. It will be found to be

$$x_1 = 0.325x_2 + 1.383x_3 - 0.383x_4$$

or, transferring the origins and expressing the equation in terms of percentage ratios,

$$X_1 = -31.1 + 0.325X_2 + 1.383X_3 - 0.383X_4$$

or, again, in terms of percentage changes (ratio -100) :

Percentage change in pauperism

$$= +1.4 \text{ per cent.}$$
$+0.325$ times the change in out-relief ratio
$+1.383$,, ,, ,, proportion of old
-0.383 ,, ,, ,, population

These results render the interpretation of the total coefficients, which might be equally consistent with several hypotheses, more clear and definite. The questions would arise, for instance, whether the correlation of changes in pauperism with changes in out-relief might not be due to correlation of the latter with the other factors introduced, and whether the negative correlation with changes in population might not be due solely to the correlation of the latter with changes in the proportion of old. As a matter of fact, the partial correlations of changes in pauperism with changes in out-relief and in proportion of old are slightly less than the total correlations, but the partial correlation with changes in population is numerically greater, the figures being:

$$r_{12} = +0.52 \qquad r_{12.34} = +0.46$$
$$r_{13} = +0.41 \qquad r_{13.24} = +0.28$$
$$r_{14} = -0.14 \qquad r_{14.23} = -0.36$$

So far, then, as we have taken the factors of the case into account, there appears to be a true correlation between changes in pauperism and changes in out-relief, proportion of old and population — the latter serving, of course, as some index to changes in general prosperity. The relative influences of the three factors are indicated by the regression equation above. (For the full discussion of the case, cf. *Jour. Roy. Stat. Soc.*, vol. 62, 1899.)

Aids to Calculation.

14.17. To facilitate the computation of partial correlation and regression coefficients, various tables of such quantities as

$$1 - r^2, \qquad \sqrt{1 - r^2}, \qquad \frac{1}{\sqrt{(1 - r_{13}^2)(1 - r_{12}^2)}}$$

have been prepared. See, for instance, refs. (610) and (611).

The Generalised Scatter Diagram.

14.18. The scatter diagram in two dimensions may be generalised to three dimensions, and may also be used as a mental construct for higher dimensions, though no actual model can of course be made.

Consider the case of three variates. The values of X_1, X_2 and X_3 associated with any given individual may be regarded as determining a point in space whose co-ordinates are X_1, X_2 and X_3. The totality of individuals will therefore give us a swarm of points in three-dimensional space, which will lie distributed in certain ways about planes of regression.

Fig. 14.1 is drawn from a model representing the data of Example 14.1.

A

B
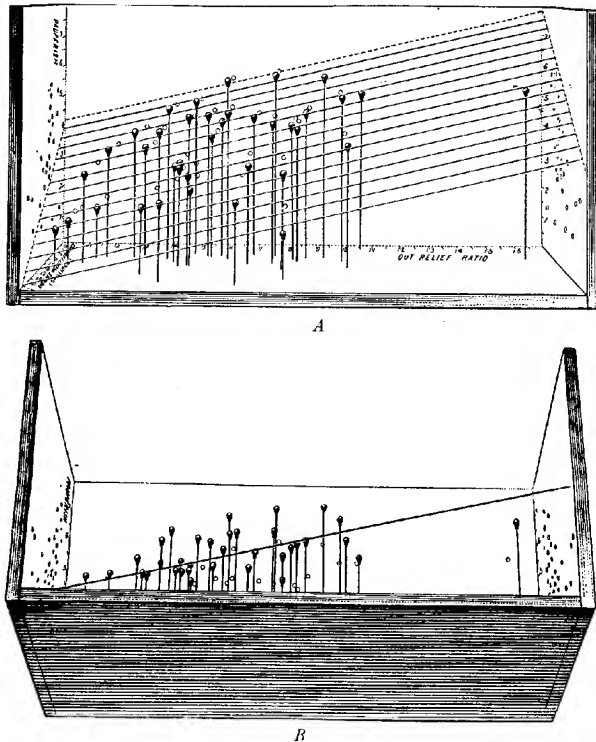
FIG. 14.1.—Model Illustrating the Correlation between Three Variables: (1) Average Weekly Earnings of Agricultural Labourers (data, Example 14.1 and Exercise 11.2); (2) Pauperism (percentage of the population in receipt of Poor Law Relief); (3) Out-relief Ratio (numbers given relief in their homes to one in the workhouse). A, front view; B, view of model tilted till the plane of regression for pauperism on the two remaining variables is seen as a straight line.

Four pieces of wood are fixed together like the bottom and three sides of a box. Supposing the open side to face the observer, a scale of pauperism is drawn vertically upwards along the left-hand angle at the back of the "box," the scale starting from zero, as very small values of pauperism occur; a scale of out-relief ratio is taken along the angle between the back

and bottom of the box, starting from zero at the left; finally, the scale of earnings is drawn out towards the observer along the angle between the left-hand side and the bottom, but as earnings lower than 12s. do not occur, the scale may start from 12s. at the corner. Suitable scales are: pauperism, 1 in. = 1 per cent.; out-relief ratio, 1 in. = 1 unit; earnings, 1 in. = 1s.; and the inside measures of the model may then be 17 in. × 10 in. × 8 in. high, the dimensions of the model constructed. Given these three scales, any set of observed values determines a point within the "box." The *earnings* and *out-relief ratio* for some one union are noted first, and the corresponding point marked on the baseboard; a steel wire is then inserted vertically in the base at this point and cut off at the height corresponding, on the scale chosen, to the pauperism in the same union, being finally capped with a small ball or knob to mark the "point" clearly. The model shows very well the general tendency of the pauperism to be the higher the lower the wages and the higher the out-relief, for the highest points lie towards the back and right-hand side of the model. If some representation of all three equations of regression were to be inserted in the model, the result would be rather confusing; so the most important equation, viz. the second, giving the average rate of pauperism in terms of the other variables, may be chosen. This equation represents a plane; the lines in which it cuts the right- and left-hand sides of the "box" should be marked, holes drilled at equal intervals on these lines on the opposite sides of the box (the holes facing each other) and threads stretched through these holes, thus outlining the plane as shown in the figure. In the actual model the correlation diagrams corresponding to the three pairs of variables were drawn on the back, sides and base: they represent, of course, the elevations and plan of the points.

The student possessing some skill in handicraft would find it worth while to make such a model for some case of interest to himself, and to study on it thoroughly the nature of the plane of regression, and the relations of the partial and total correlations.

Coefficient of Multiple Correlation.

14.19. Consider the regression equation for x_1,

$$x_1 = b_{12.3 \ldots n} x_2 + b_{13.2 \ldots n} x_3 + \ldots + b_{1n.2 \ldots (n-1)} x_n$$

Let us write the right-hand side of this equation as $e_{1.23 \ldots n}$, so that in virtue of (14.2),

$$e_{1.23 \ldots n} = x_1 - x_{1.23 \ldots n} \qquad \cdot \qquad \cdot \qquad (14.14)$$

Now consider the correlation between x_1 and $e_{1.23 \ldots n}$. We have in virtue of the theorem of **14.10**:

$$
\begin{aligned}
S(x_1 e_{1.23 \ldots n}) &= S\{x_1(x_1 - x_{1.23 \ldots n})\} \\
&= S(x_1{}^2) - S\{x_1(x_{1.23 \ldots n})\} \\
&= S(x_1{}^2) - S(x_{1.23 \ldots n})^2 \\
&= N(\sigma_1^2 - \sigma_{1.23 \ldots n}^2)
\end{aligned}
$$

Also,

$$
\begin{aligned}
S(e_{1.23 \ldots n})^2 &= S(x_1 - x_{1.23 \ldots n})^2 \\
&= N(\sigma_1^2 - \sigma_{1.23 \ldots n}^2)
\end{aligned}
$$

Hence, the correlation between x_1 and $e_{1.23 \ldots n}$

$$= \frac{\sigma_1^2 - \sigma_{1.23 \ldots n}^2}{\sigma_1 \sqrt{\sigma_1^2 - \sigma_{1.23 \ldots n}^2}}$$

$$= \frac{\sqrt{\sigma_1^2 - \sigma_{1.23 \ldots n}^2}}{\sigma_1}$$

We shall call this quantity $R_{1(23 \ldots n)}$. We have immediately:

$$\sigma_{1.23 \ldots n}^2 = \sigma_1^2 (1 - R_{1(23 \ldots n)}^2) . \qquad . \qquad . \quad (14.15)$$

$R_{1(2 \ldots n)}$ is called the **multiple correlation coefficient** between x_1 and $x_2 \ldots x_n$. We have, similarly, multiple correlations between x_1 and fewer variables. $R_{1(2 \ldots n)}$ is called an $(n-1)$-fold multiple correlation coefficient. $R_{1(2 \ldots \overline{n-1})}$ would be an $(n-2)$-fold coefficient, and so on.

14.20. The value of R may be calculated either directly from equation (14.15), or by substituting in that equation the value of $\sigma_{1.23 \ldots n}^2$ obtained in (14.10), which gives:

$$1 - R_{1(23 \ldots n)}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2) \; . \; . \; . \; (1 - r_{1n.23 \ldots (n-1)}^2) \quad (14.16)$$

Properties of the Multiple Correlation Coefficient.

14.21. $R_{1(23 \ldots n)}$, being the correlation between x_1 and $e_{1.23 \ldots n}$, measures how closely x_1 can be represented by the regression equation. If $R = 1$, x_1 can be perfectly represented by such an equation, *i.e.* is a linear function of $x_2 \ldots x_n$. In this case $\sigma_{1.23 \ldots n}^2 = 0$, *i.e.* all the residuals are zero.

It may, in fact, be shown that $R_{1(23 \ldots n)}$ is greater than the correlation between x_1 and any linear function of $x_2 \ldots x_n$ other than that expressed in the regression equation, *i.e.* $e_{1.23 \ldots n}$. Putting this another way, the regression coefficients in $e_{1.23 \ldots n}$ may be determined by the condition that the correlation between x_1 and $e_{1.23 \ldots n}$ is a maximum.

R is Necessarily Positive or Zero.

14.22. This is true, since the product term $S(x_1 e_{1.23 \ldots n})$ is positive, being equal to $N(\sigma_1^2 - \sigma_{1.23 \ldots n}^2)$, and we see from (14.10) that $\sigma_1^2 \geqslant \sigma_{1.23 \ldots n}^2$.

Further, from   4.16),

$$1 - R_{1(23 \ldots n)}^2 \leqslant 1 - r_{12}^2$$

i.e. R is not numerically less than r_{12}. Similarly, it is not numerically less than any other total or partial correlation coefficient which can appear in (14.16). Hence, $R_{1(2 \ldots n)}$ *is not numerically less than any possible constituent coefficient of correlation.*

It follows from this that if $R_{1(2 \ldots n)} = 0$, all the correlation coefficients involving x_1 are zero, *i.e.* the variate x_1 is *completely uncorrelated with the other variates.*

14.23. Further, even if all the variables $X_1, X_2, \ldots X_n$ were strictly uncorrelated in the original universe as a whole, we should expect $r_{12}, r_{13.2}, r_{14.23}$, etc. to exhibit values (whether positive or negative) differing

from zero in a limited sample. Hence, R will not tend, on an average of such samples, to be zero, but will fluctuate round some mean value. This mean value will be the greater the smaller the number of observations in the sample, and also the greater the number of variables. When only a small number of observations is available it is, accordingly, little use to deal with a large number of variables. As a limiting case, it is evident that if we deal with n variables and possess only n observations, all the partial correlations of the highest possible order will be unity. We shall deal with the question of the significance of an observed value of R in a later chapter (23.45).

Example 14.3.—In Example 14.1 we found :

$$r_{12} = -0.66$$
$$r_{13.2} = +0.44 \quad _,$$

Hence, from (14.16),

$$1 - R^2_{1(23)} = \{1 - (0.66)^2\}\{1 - (0.44)^2\}$$
$$= 0.455$$

whence

$$R_{1(23)} = 0.74$$

Similarly, it will be found that

$$R_{2(13)} = 0.84$$

and

$$R_{3(12)} = 0.70$$

The student may verify by inspection that these values are greater than the corresponding constituent values.

Expression of Regressions and Correlations in terms of Coefficients of Higher Orders.

14.24. It is obvious that as equations (14.12) and (14.13) enable us to express regressions and correlations of higher orders in terms of those of lower orders, we must similarly be able to express the coefficients of lower in terms of those of higher orders. Such expressions are sometimes useful for theoretical work. Using the same method of expansion as in previous cases, we have :

$$0 = S(x_{1.23 \, \ldots \, n} x_{2.34 \, \ldots \, (n-1)})$$
$$= S(x_1 x_{2.34 \, \ldots \, (n-1)}) - b_{12.34 \, \ldots \, n} S(x_2 x_{2.34 \, \ldots \, (n-1)})$$
$$- b_{1n.23 \, \ldots \, (n-1)} S(x_n x_{2.34 \, \ldots \, (n-1)})$$

That is,

$$b_{12.34 \, \ldots \, (n-1)} = b_{12.34 \, \ldots \, n} + b_{1n.23 \, \ldots \, (n-1)} b_{n2.34 \, \ldots \, (n-1)}$$

In this equation the coefficient on the left and the last on the right are of order $n-3$, the other two of order $n-2$. We therefore wish to eliminate the last coefficient on the right. Interchanging the suffixes 1 for n and n for 1, we have :

$$b_{n2.34 \, \ldots \, (n-1)} = b_{n2.13 \, \ldots \, (n-1)} + b_{n1.23 \, \ldots \, (n-1)} b_{12.34 \, \ldots \, (n-1)}$$

Substituting this value for $b_{n2.34 \ldots (n-1)}$ in the first equation, we have:

$$b_{12.34 \ldots (n-1)} = \frac{b_{12.34 \ldots n} + b_{1n.23 \ldots (n-1)}b_{n2.13 \ldots (n-1)}}{1 - b_{1n.23 \ldots (n-1)}b_{n1.23 \ldots (n-1)}} \quad (14.17)$$

This is the required equation for the regressions; it is the equation

$$b_{12} = \frac{b_{12.n} + b_{1n.2}b_{n2.1}}{1 - b_{1n.2}b_{n1.2}}$$

with secondary suffixes $34 \ldots (n-1)$ added throughout. The corresponding equation for the correlations is obtained at once by writing down equation (14.17) for $b_{21.34 \ldots (n-1)}$ and taking the square root of the product; this gives:

$$r_{12.34 \ldots (n-1)} = \frac{r_{12.34 \ldots n} + r_{1n.23 \ldots (n-1)}r_{2n.13 \ldots (n-1)}}{(1 - r_{1n.23 \ldots (n-1)}^2)^{\frac{1}{2}}(1 - r_{2n.13 \ldots (n-1)}^2)^{\frac{1}{2}}} \quad (14.18)$$

which is similarly the equation

$$r_{12} = \frac{r_{12.n} + r_{1n.2}r_{2n.1}}{(1 - r_{1n.2}^2)^{\frac{1}{2}}(1 - r_{2n.1}^2)^{\frac{1}{2}}}$$

with the secondary suffixes $34 \ldots (n-1)$ added throughout.

Conditions of Consistence among Correlation Coefficients.

14.25. Equations (14.13) and (14.18) imply that certain limiting inequalities must hold between the correlation coefficients in the expression on the right in each case in order that real values (values between ± 1) may be obtained for the correlation coefficient on the left. These inequalities correspond precisely with those " conditions of consistence " between class-frequencies with which we dealt in Chapter 2, but we propose to treat them only briefly here. Writing (14.13) in its simplest form for $r_{12.3}$, we must have $r_{12.3}^2 \leqslant 1$ or

$$\frac{(r_{12} - r_{13}r_{23})^2}{(1 - r_{13}^2)(1 - r_{23}^2)} \leqslant 1$$

that is,

$$r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23} \leqslant 1 \quad . \quad . \quad . \quad (14.19)$$

if the three r's are consistent with one another. If we take r_{12}, r_{13} as known, this gives as limits for r_{23},

$$r_{12}r_{13} \pm \sqrt{1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2}$$

Similarly, writing (14.18) in its simplest form for r_{12} in terms of $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$, we must have:

$$r_{12.3}^2 + r_{13.2}^2 + r_{23.1}^2 + 2r_{12.3}r_{13.2}r_{23.1} \leqslant 1 \quad . \quad . \quad (14.20)$$

and therefore, if $r_{12.3}$ and $r_{13.2}$ are given, $r_{23.1}$ must lie between the limits

$$-r_{12.3}r_{13.2} \pm \sqrt{1 - r_{12.3}^2 - r_{13.2}^2 + r_{12.3}^2 r_{13.2}^2} \quad .$$

The following table gives the limits of the third coefficient, in a few special cases, for the three coefficients of zero order and of the first order respectively :—

Value of		Limits of	
r_{12} or $r_{12.3}$	r_{13} or $r_{13.2}$	r_{23}	$r_{23.1}$
0	0	±1	±1
±1	±1	+1	−1
±1	∓1	−1	+1
±√0·5	±√0·5	0, +1	0, −1
±√0·5	∓√0·5	0, −1	0, +1

The student should notice that the set of three coefficients of order zero and value unity are only consistent if either one only, or all three, are positive, *i.e.* +1, +1, +1, or −1, −1, +1 ; but not −1, −1, −1. On the other hand, the set of three coefficients of the first order and value unity are only consistent if one only, or all three, are negative : the only consistent sets are +1, +1, −1 and −1, −1, −1. The values of the two given r's need to be very high if even the sign of the third can be inferred ; if the two are equal, they must be at least equal to $\sqrt{0·5}$ or 0·707 . . . Finally, it may be noted that no two values for the known coefficients ever permit an inference of the value zero for the third ; the fact that 1 and 2, 1 and 3 are uncorrelated, pair and pair, permits no inference of any kind as to the correlation between 2 and 3, which may lie anywhere between +1 and −1.

Fallacies in the Interpretation of Correlation Coefficients.

14.26. We do not think it necessary to add to this chapter a detailed discussion of the nature of fallacies on which the theory of multiple correlation throws much light. The general nature of such fallacies is the same as for the case of attributes, and was discussed fully in Chapter 4. It suffices to point out the principal sources of fallacy which are suggested at once by the form of the partial correlation

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \qquad . \qquad . \qquad . \qquad (a)$$

and from the form of the corresponding expression for r_{12} in terms of the partial coefficients :

$$r_{12} = \frac{r_{12.3} + r_{13.2}r_{23.1}}{\sqrt{(1 - r_{13.2}^2)(1 - r_{23.1}^2)}} \qquad . \qquad . \qquad . \qquad (b)$$

From the form of the numerator of (a) it is evident (1) that even if r_{12} be zero, $r_{12.3}$ will not be zero unless either r_{13} or r_{23}, or both, are zero. If r_{13} and r_{23} are of the same sign, the partial correlation will be negative ; if of opposite sign, positive. Thus the quantity of a crop might appear to be unaffected, say, by the amount of rainfall during some period preceding harvest : this might be due merely to a correlation between rain and low temperature, the partial correlation between crop and rainfall being

positive and important. We may thus easily misinterpret a coefficient of correlation which is zero. (2) $r_{12.3}$ may be, indeed often is, of opposite sign to r_{12}, and this may lead to still more serious errors of interpretation.

From the form of the numerator of (b), on the other hand, we see that, conversely, r_{12} will not be zero even though $r_{12.3}$ is zero, unless either $r_{13.2}$ or $r_{23.1}$ is zero. This corresponds to the theorem of **4.12**, and indicates a source of fallacies similar to those there discussed.

14.27. We have seen that $r_{12.3}$ is the correlation between $x_{1.3}$ and $x_{2.3}$, and that we might determine the value of this partial correlation by drawing up the actual correlation table for the two residuals in question. Suppose, however, that instead of drawing up a single table we drew up a series of tables for values of $x_{1.3}$ and $x_{2.3}$ associated with values of x_3 lying within successive class-intervals of its range. In general, the value of $r_{12.3}$ would not be the same (or approximately the same) for all such tables, but would exhibit some systematic change as the value of x_3 increased. Hence $r_{12.3}$ should be regarded, in general, as of the nature of an average correlation : the cases in which it measures the correlation between $x_{1.3}$ and $x_{2.3}$ for *every* value of x_3 (*cf.* below, **14.31**) are probably exceptional. The process for determining partial associations (*cf.* Chapter 4) is, it will be remembered, thorough and complete, as we always obtain the actual tables exhibiting the association between, say, A and B in the universe of C's and the universe of γ's : that two such associations may differ materially is illustrated by Example 4.1, page 52. It might sometimes serve as a useful check on partial correlation work to reclassify the observations by the fundamental methods of Chapter 4. For the general case an extension of the method of the " correlation ratio " (**13.5**) might be useful, though exceedingly laborious.

Multivariate Normal Correlation.

14.28. The theorems and results of Chapter 12 in regard to normal correlation can be extended to the case of n variates, which we have studied in this chapter.

In fact, suppose we have n variates $x_1, x_2, x_3, \ldots x_n$, measured from their respective means, with standard deviations $\sigma_1, \sigma_2, \sigma_3, \ldots \sigma_n$. Let us first consider the simple case in which they are normally distributed and each is completely independent of the others.

Then, if $y_1 \ldots {}_n$ denote the frequency of the combination of deviations $x_1, x_2, \ldots x_n$, we have:

$$y_{12} \ldots {}_n = y'_{12} \ldots {}_n e^{-\frac{1}{2}\phi(x_1, x_2, \ldots x_n)}$$

where

$$\phi(x_1, x_2, \ldots x_n) = \frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} + \frac{x_3^2}{\sigma_3^2} + \ldots + \frac{x_n^2}{\sigma_n^2} \qquad . \quad (14.21)$$

Now consider the variates $x_1, x_{2.1}, x_{3.12}, \ldots x_{n.12} \ldots {}_{(n-1)}$. Whether $x_1, x_2, \ldots x_n$ are correlated or not, these variates are uncorrelated, in virtue of **14.10**. Let us further suppose they are independent and normally distributed. Then their distribution is given by

$$y_{12} \ldots {}_n = y'_{12} \ldots {}_n e^{-\frac{1}{2}\phi(x_1, x_{2.1}, \ldots x_{n.12} \ldots {}_{(n-1)})} \qquad . \quad (14.22)$$

where

$$\phi(x_1, x_{2.1}, \ldots x_{n.12 \ldots (n-1)}) = \frac{x_1^2}{\sigma_1^2} + \frac{x_{2.1}^2}{\sigma_{2.1}^2} + \ldots + \frac{x_{n.12 \ldots (n-1)}^2}{\sigma_{n.12 \ldots (n-1)}^2} \quad (14.23)$$

and

$$y_{12 \ldots n}' = \frac{N}{(2\pi)^{\frac{n}{2}} \sigma_1 \sigma_{2.1} \ldots \sigma_{n.12 \ldots (n-1)}} \cdot \qquad (14.24)$$

The expression (14.23) may be put in a more convenient form. It may be shown, but we omit the proof, that

$$\phi = \frac{x_1^2}{\sigma_{1.23 \ldots n}^2} + \frac{x_2^2}{\sigma_{2.13 \ldots n}^2} + \ldots + \frac{x_n^2}{\sigma_{n.12 \ldots (n-1)}^2} \qquad (14.25)$$

$$-2r_{12.3 \ldots n}\frac{x_1 x_2}{\sigma_{1.23 \ldots n}\sigma_{3.13 \ldots n}} - \ldots - 2r_{(n-1)n.12 \ldots (n-2)}\frac{x_{n-1}x_n}{\sigma_{n-1.1 \ldots (n-2)n}\sigma_{n.1 \ldots (n-1)}}$$

which exhibits the form as symmetrical in $x_1 \ldots x_n$.

Now, we showed in **14.13** that

$$\sigma_{1.23 \ldots n}^2 = \frac{\omega}{\omega_{11}}\sigma_1^2$$

etc.

In precisely the same way it may be shown that

$$\sigma_{1.23 \ldots n}\sigma_{2.13 \ldots n}/r_{12.3 \ldots n} = -\frac{\omega}{\omega_{12}}\sigma_1\sigma_2$$

ω_{12} being the minor in ω of the term in the first row and the second column.

If we substitute these and analogous values in (14.22), we get:

$$y_{12 \ldots n} = \frac{N}{(2\pi)^{\frac{n}{2}}\sigma_1\sigma_2 \ldots \sigma_n \sqrt{\omega}}e^{-\frac{1}{2}\phi}$$

where

$$\phi = \frac{1}{\omega}\left\{\omega_{11}\frac{x_1^2}{\sigma_1^2} + \omega_{22}\frac{x_2^2}{\sigma_2^2} + \ldots + 2\omega_{12}\frac{x_1 x_2}{\sigma_1\sigma_2} + \ldots + 2\omega_{n,n-1}\frac{x_n x_{n-1}}{\sigma_n\sigma_{n-1}}\right\} \quad (14.26)$$

This is a form which is very frequently quoted.

14.29. From these formulæ several important results follow immediately.

In the first place, for any fixed values $h_2 \ldots h_n$ of $x_2 \ldots x_n$, the exponent (14.25) becomes:

$$\frac{x_1^2}{\sigma_{1.23 \ldots n}^2} - 2r_{12.34 \ldots n}\frac{x_1 h_2}{\sigma_{1.23 \ldots n}\sigma_{2.13 \ldots n}} - \ldots - \frac{2r_{1n.2 \ldots (n-1)}x_1 h_n}{\sigma_{1.23 \ldots n}\sigma_{n.1 \ldots (n-1)}} + \text{constant terms}$$

$$= \left\{\frac{x_1}{\sigma_{1.23 \ldots n}} - \frac{r_{12.3 \ldots n}h_2}{\sigma_{2.13 \ldots n}} - \ldots - \frac{r_{1n.3 \ldots (n-1)}h_n}{\sigma_{n.1 \ldots (n-1)}}\right\}^2 + \text{constant terms}$$

Hence x_1 is distributed normally about the mean, $\dot{m}_1$, given by

$$\frac{m_1}{\sigma_{1.23\ldots n}} = \frac{r_{12.3\ldots n}}{\sigma_{2.13\ldots n}}h_2 + \ldots + \frac{r_{1n.2\ldots(n-1)}}{\sigma_{n.1\ldots(n-1)}}h_n \quad . \quad (14.27)$$

Hence every array of every order is normally distributed.

It follows in a similar way that any linear function of the x's is distributed normally.

In particular, all deviations of any order and with any number of suffixes are normally distributed.

14.30. Secondly, as will be seen from (14.27), the regression of x_1 on the other variables is linear. It follows that the regression of any variate on any or all of the others is linear. In (14.27), for instance, the expressions $\dfrac{r_{12.3\ldots n}\sigma_{1.23\ldots n}}{\sigma_{2.13\ldots n}}$, etc., are the partial regressions $b_{12.3\ldots n}$, etc.

14.31. Finally if, in equation (14.23), any fixed values be assigned to $x_{3.12}$ and all the following deviations, the correlation between x_1 and x_2, on expanding $x_{2.1}$, is, as we have seen, normal correlation. Similarly, if any fixed values be assigned to x_1, to $x_{4.123}$, and all the following deviations, on reducing $x_{3.12}$ to the second order we shall find that the correlation between $x_{2.1}$ and $x_{3.1}$ is normal correlation, the correlation coefficient being $r_{23.1}$, and so on. That is to say, using k to denote any group of secondary suffixes, (1) *the correlation between any two deviations $x_{m.k}$ and $x_{n.k}$ is normal correlation;* (2) *the correlation between the said deviations is $r_{mn.k}$ whatever the particular fixed values assigned to the remaining deviations.* The latter conclusion, it will be seen, renders the meaning of partial correlation coefficients much more definite in the case of normal correlation than in the general case. In the general case $r_{mn.k}$ represents merely the average correlation, so to speak, between $x_{m.k}$ and $x_{n.k}$: in the normal case $r_{mn.k}$ is constant for all the subgroups corresponding to particular assigned values of the other variables. Thus in the case of three variables which are normally correlated, if we assign any given value to x_3, the correlation between the associated values of x_1 and x_2 is $r_{12.3}$: in the general case $r_{12.3}$, if actually worked out for the various sub-groups corresponding, say, to increasing values of x_3, would probably exhibit some continuous change, increasing or decreasing as the case might be.

SUMMARY.

1. The regression equation of x_1 on x_2, x_3, . . . x_n is written:

$$\dot{x}_1 = b_{12.34\ldots n}x_2 + b_{13.24\ldots n}x_3 + \ldots + b_{1n.23\ldots(n-1)}x_n$$

The deviation $x_{1.23\ldots n}$ is defined as

$$x_1 - b_{12.34\ldots n}x_2 - b_{13.24\ldots n}x_3 - \ldots - b_{1n.23\ldots(n-1)}x_n$$

and $\sigma_{1.23\ldots n}$ is the standard deviation of $x_{1.23\ldots n}$.

2. The equations giving the regression coefficients are:

$$S(x_2 x_{1.23 \ldots n}) = 0$$
$$S(x_3 x_{1.23 \ldots n}) = 0$$
$$\cdot \qquad \cdot \qquad \cdot$$
$$S(x_n x_{1.23 \ldots n}) = 0$$

and similar equations with $x_{2.13 \ldots n}$, etc.

3. The product-sum of any two deviations is unaltered by omitting any or all of the secondary subscripts of either which are common to the two; conversely, the product-sum of any deviation of order p with a deviation of order $p + q$, the p subscripts being the same in each case, is unaltered by adding to the secondary subscripts of the former any or all of the q additional subscripts of the latter.

4. $$b_{12.34 \ldots n} = r_{12.34 \ldots n} \frac{\sigma_{1.34 \ldots n}}{\sigma_{2.34 \ldots n}}$$

5. Any standard deviation of order p can be expressed in terms of a standard deviation of order $p - 1$ and a correlation of order $p - 1$. In fact,

$$\sigma_{1.23 \ldots n}^2 = \sigma_{1.23 \ldots (n-1)}^2 (1 - r_{1n.23 \ldots (n-1)}^2)$$

6. $$\sigma_{p.23 \ldots n}^2 = \frac{\omega \sigma_p^2}{\omega_{pp}}$$

where ω is the determinant

$$\begin{vmatrix} 1 & r_{12} & r_{13} & \ldots & r_{1n} \\ r_{21} & 1 & r_{23} & \ldots & r_{2n} \\ \cdot & \cdot & & \cdot & \\ r_{n1} & r_{n2} & r_{n3} & \ldots & 1 \end{vmatrix}$$

and ω_{pp} is the minor of the element in the pth row and the pth column.

7. Any regression of order p may be expressed in terms of regressions of order $p - 1$. In fact,

$$b_{12.34 \ldots n} = \frac{b_{12.34 \ldots (n-1)} - b_{1n.34 \ldots (n-1)} b_{n2.34 \ldots (n-1)}}{1 - b_{2n.34 \ldots (n-1)} b_{n2.34 \ldots (n-1)}}$$

8. Similarly, for correlations :

$$r_{12.34 \ldots n} = \frac{r_{12.34 \ldots (n-1)} - r_{1n.34 \ldots (n-1)} r_{2n.34 \ldots (n-1)}}{(1 - r_{1n.34 \ldots (n-1)}^2)^{\frac{1}{2}} (1 - r_{2n.34 \ldots (n-1)}^2)^{\frac{1}{2}}}$$

9. The coefficient of multiple correlation $R_{1(23 \ldots n)}$ is given by

$$\sigma_{1.23 \ldots n}^2 = \sigma_1^2 (1 - R_{1(23 \ldots n)}^2)$$

or

$$\frac{\omega}{\omega_{11}} = 1 - R_{1(23 \ldots n)}^2$$

Also,

$$1 - R_{1(23 \ldots n)}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2) \ldots (1 - r_{1n.23 \ldots (n-1)}^2)$$

10. R is necessarily not less than zero. If it is zero, the variate to which it refers is completely uncorrelated with the other variates. If $R = 1$, there is a linear relation between the variates.

11. The multivariate normal surface may be written :

$$y_{12\ \ldots\ n} = \frac{N}{(2\pi)^{\frac{n}{2}}\sigma_1\sigma_2\ \ldots\ \sigma_n\sqrt{\omega}}\,e^{-\frac{1}{2}\phi}$$

where

$$\phi = \frac{1}{\omega}\left\{\omega_{11}\frac{x_1^2}{\sigma_1^2} + \omega_{22}\frac{x_2^2}{\sigma_2^2} + \ \ldots\ + 2\omega_{12}\frac{x_1x_2}{\sigma_1\sigma_2} + \ \ldots\ + 2\omega_{n.n-1}\frac{x_nx_{n-1}}{\sigma_n\sigma_{n-1}}\right\}$$

EXERCISES.

14.1. (Ref. (299).) The following means, standard deviations and correlations are found for

$X_1 =$ Seed-hay crops in cwts. per acre,
$X_2 =$ Spring rainfall in inches,
$X_3 =$ Accumulated temperature above $42°$ F. in spring,

in a certain district of England during twenty years.

$M_1 = 28 \cdot 02$	$\sigma_1 = 4 \cdot 42$	$r_{12} = +0 \cdot 80$
$M_2 = 4 \cdot 91$	$\sigma_2 = 1 \cdot 10$	$r_{13} = -0 \cdot 40$
$M_3 = 594$	$\sigma_3 = 85$	$r_{23} = -0 \cdot 56$

Find the partial correlations and the regression equation for hay-crop on spring rainfall and accumulated temperature.

14.2. In Exercise 14.1, find the multiple correlation coefficient of each variate on the other two.

14.3. (The following figures must be taken as an illustration only: the data on which they were based do not refer to uniform times or areas.)

$X_1 =$ Deaths of infants under 1 year per 1000 births in same year (infantile mortality).
$X_2 =$ Number per thousand of married women occupied for gain.
$X_3 =$ Death-rate of persons over 5 years of age per 10,000.
$X_4 =$ Number per thousand of population living two or more to a room (overcrowding).

Taking the figures below for thirty urban areas in England and Wales, find the partial correlations and the regression equation for infantile mortality on the other factors.

$M_1 = 164$	$\sigma_1 = 20 \cdot 0$	$r_{12} = +0 \cdot 49$	$r_{23} = +0 \cdot 15$
$M_2 = 158$	$\sigma_2 = 74 \cdot 9$	$r_{13} = +0 \cdot 78$	$r_{24} = -0 \cdot 37$
$M_3 = 143$	$\sigma_3 = 22 \cdot 4$	$r_{14} = +0 \cdot 20$	$r_{34} = +0 \cdot 23$
$M_4 = 205$	$\sigma_1 = 130 \cdot 0$		

14.4. In Exercise 14.3, find the multiple correlation coefficient of X_1 on X_2 and X_3; and of X_1 on the other three variates.

14.5. (Data from W. F. Ogburn, "Factors in the Variation of Crime among Cities," *Jour. Amer. Stat. Assoc.*, vol. 30, 1935, pp. 12–34.)

For certain large cities in the U.S.A. :

$X_1 =$ Crime rate, being the number of known offences per thousand of population.
$X_2 =$ Percentage of male inhabitants.
$X_3 =$ Percentage of total inhabitants who are foreign-born males.

X_4 = Number of children under 5 years of age per thousand married women between 15 and 44 years of age.

X_5 = Church membership, being number of church members 13 years of age and over per 100 of total population 13 years of age and over.

$M_1 =$ 19·9	$\sigma_1 =$ 7·9	$r_{12} = +0·44$	$r_{24} = -0·19$
$M_2 =$ 49·2	$\sigma_2 =$ 1·3	$r_{13} = -0·34$	$r_{25} = -0·35$
$M_3 =$ 10·2	$\sigma_3 =$ 4·6	$r_{14} = -0·31$	$r_{34} = +0·44$
$M_4 =$ 481·4	$\sigma_4 =$ 74·4	$r_{15} = -0·14$	$r_{35} = +0·33$
$M_5 =$ 41·6	$\sigma_5 =$ 10·8	$r_{23} = +0·25$	$r_{45} = +0·85$

Find the regression equation of X_1 on the other four variables. Find also $R_{1(2345)}$.

Find, further, $r_{15.3}$, $r_{15.4}$ and $r_{15.34}$. Discuss the influence of church membership on crime for these data.

14.6. Show that for n variates there are $^nC_2$ total correlation coefficients, $(n-2)^nC_2$ correlation coefficients of order 1, $^{n-2}C_2{}^nC_2$ correlation coefficients of order 2, and $^{n-2}C_s{}^nC_2$ of order s. Hence show that there are $n(n-1)2^{n-3}$ correlation coefficients and $n(n-1)2^{n-2}$ regression coefficients.

14.7. Find the number of multiple correlation coefficients of order s and the total number of such coefficients for n variables.

14.8. If all the correlations of order zero are equal, say $= r$, what are the values of the partial correlations of successive orders?

Under the same conditions, what is the limiting value of r if all the equal correlations are negative and n variables have been observed?

14.9. Write down from inspection the values of the partial correlations for the three variables

$$X_1, X_2 \text{ and } X_3 = aX_1 + bX_2$$

14.10. If the relation

$$ax_1 + bx_2 + cx_3 = 0$$

holds for all sets of values of x_1, x_2 and x_3, what must the partial correlations be?

CORRELATION: ILLUSTRATIONS AND PRACTICAL METHODS.

15.1. The student—especially the student of economic statistics, to whom this chapter is principally addressed—should be careful to note that the coefficient of correlation, like an average or a measure of dispersion, only exhibits in a summary and comprehensible form one particular aspect of the facts on which it is based, and the real difficulties arise in the interpretation of the coefficient when obtained. The value of the coefficient may be consistent with some given hypothesis, but it may be equally consistent with others ; and not only are care and judgment essential for the discussion of such possible hypotheses, but also a thorough knowledge of the facts in all other possible aspects. Further, care should be exercised from the commencement in the selection of the variables between which the correlation shall be determined. The variables should be defined in such a way as to render the correlations as readily interpretable as possible, and, if several are to be dealt with, they should afford the answers to specific and definite questions. Unfortunately, the field of choice is frequently very much limited, by deficiencies in the available data and so forth, and consequently practical possibilities as well as ideal requirements have to be taken into account. No general rules can be laid down, but the following are given as illustrations of the sort of points that have to be considered.

15.2. *Example 15.1.*—It is required to throw some light on the variations of pauperism in the unions (unions of parishes) of England. (*Cf.* Yule, ref. (334)—investigation carried out in 1898.)

On the whole, it would seem best to correlate *changes* in pauperism with *changes* in various possible factors. If we say that a high rate of pauperism in some district is due to lax administration, we presumably mean that as administration became lax, pauperism rose, or that if administration were more strict, pauperism would decrease ; if we say that the high pauperism is due to the depressed condition of industry, we mean that when industry recovers pauperism will fall. When we say, in fact, that any one variable is a factor of pauperism, we mean that *changes* in that variable are accompanied by *changes* in the percentage of the population in receipt of relief, either in the same or the reverse direction. It will be better, therefore, to deal with changes in pauperism and possible factors. The next question is what factors to choose.

15.3. The possible factors may be grouped under three heads :

(a) *Administration.*—Changes in the method or strictness of administration of the law.

(b) *Environment.*—Changes in economic conditions (wages, prices, employment), social conditions (residential or industrial character of the

district, density of population, nationality of population) or moral conditions (as illustrated, *e.g.*, by the statistics of crime).

(c) *Age Distribution.*—The percentage of the population between given age-limits in receipt of relief increases very rapidly with old age, the actual figures given by one of the only two then existing returns of the age of paupers being: 2 per cent. under age 16, 1 per cent. over 16 but under 65, 20 per cent. over 65. (Return 36, 1890.)

It is practically impossible to deal with more than three factors, one from each of the above groups, or four variables altogether, including the pauperism itself. What shall we take, then, as representative variables, and how shall we best measure " pauperism " ?

15.4. *Pauperism.*—The returns give (*a*) cost, (*b*) numbers relieved. It seems better to deal with (*b*), as numbers are more important than cost from the standpoint of the moral effect of relief on the population. The returns, however, generally include both lunatics and vagrants in the totals of persons relieved ; and as the administrative methods of dealing with these two classes differ entirely from the methods applicable to ordinary pauperism, it seems better to alter the official total by excluding them. Returns are available giving the numbers in receipt of relief on 1st January and 1st July; there does not seem to be any special reason for taking the one return rather than the other, but the return for 1st January was actually used. The percentage of the population in receipt of relief on 1st January 1871, 1881 and 1891 (the three census years), less lunatics and vagrants, was therefore tabulated for each union.

15.5. *Administration.*—The most important point here, and one that lends itself readily to statistical treatment, is the relative proportion of indoor and outdoor relief (relief in the workhouse and relief in the applicant's home). The first question is, again, shall we measure this proportion by cost or by numbers ? The latter seems, as before, the simpler and more important ratio for the present purpose, though some writers have preferred the statement in terms of expenditure (*e.g.* Charles Booth, "*Aged Poor—Condition, 1894* "). If we decide on the statement in terms of numbers, we still have the choice of expressing the proportion (1) as the ratio of numbers given out-relief to numbers in the workhouse, or (2) as the percentage of numbers given out-relief on the total number relieved. The former method was chosen, partly on the simple ground that it had already been used in an earlier investigation, partly on the ground that the use of the ratio separates the higher proportions of out-relief more clearly from each other, and these differences seem to have significance. Thus a union with a ratio of 15 outdoor paupers to 1 indoor seems to be materially different from one with a ratio of, say, 10 to 1 ; but if we take, instead of the ratios, the percentages of outdoor to total paupers, the figures are 94 per cent. and 91 per cent. respectively, which are so close that they will probably fall into the same array. The ratio of numbers in receipt of outdoor relief to the numbers in the workhouse, in every union, was therefore tabulated for 1st January in the census years 1871, 1881, 1891.

15.6. *Environment.*—This is the most difficult factor of all to deal with. In Booth's work the factors tabulated were (1) persons per acre ; (2) percentage of population living two or more to a room, *i.e.* "overcrowding " ; (3) rateable value per head ("*Aged Poor—Condition* "). The data relating to overcrowding were first collected at the census of 1891, and are not

19

available for earlier years. Some trial was made of rateable value per head, but with not very satisfactory results. For any given year, and for a group of unions of somewhat similar character, *e.g.* rural, the rateable value per head appears to be highly (negatively) correlated with the pauperism, but changes in the two are not very highly correlated : probably the movements of assessments are sluggish and irregular, especially in the case of falling assessments in rural unions, and do not correspond at all accurately with the real changes in the value of agricultural land. After some consideration, it was decided to use a very simple index to the changing fortunes of a district, viz. the movement of the population itself. If the population of a district is increasing at a rate above the average, this is *prima facie* evidence that its industries are prospering ; if the population is decreasing, or not increasing as fast as the average, this strongly suggests that the industries are suffering from a temporary lack of prosperity or permanent decay. The population of every union was therefore tabulated for the censuses of 1871, 1881, 1891.

15.7. *Age Distribution.*—As already stated, the figures that are known clearly indicate a very rapid rise of the percentage relieved after 65 years of age. The percentage of the population over 65 years of age was therefore worked out for every union and tabulated from the same three censuses. This is not, of course, at all a complete index to the composition of the population as affecting the rate of pauperism, which is sensibly dependent on the proportion of the two sexes, and the numbers of children as well. As the percentage in receipt of relief was, however, 20 per cent. for those over 65, and only 1 to 2 per cent. for those under that age, it is evidently a most important index. (A more complete method might have been used by correcting the observed rate of pauperism to the basis of a standard population with given numbers of each age and sex (*cf.* Chap. 16, pages 305–306).)

15.8. The changes in each of the four quantities that had been tabulated for every union were then measured by working out the ratios for the intercensal decades 1871–81 and 1881–91, taking the value in the earlier year as 100 in each case. The percentage ratios so obtained were taken as the four variables. Further, as the conditions are and were very different for rural and for urban unions, it seemed very desirable to separate the unions into groups according to their character. But this cannot be done with any exactness : the majority of unions are of a mixed character, consisting, say, of a small town with a considerable extent of the surrounding country. It might seem best to base the classification on returns of occupations, *e.g.* the proportions of the population engaged in agriculture, but the statistics of occupations are not given in the census for individual unions. Finally, it was decided to use a classification by density of population, the grouping used being—Rural, 0·3 person per acre or less ; Mixed, more than 0·3 but not more than 1 person per acre ; Urban, more than 1 person per acre. The metropolitan unions were also treated by themselves. The limit 0·3 for rural unions was suggested by the density of those agricultural unions the conditions in which were investigated by the Labour Commission which reported in 1894 : the average density of these was 0·25, and 34 of the 38 were under 0·3. The lower limit of density for urban unions—1 per acre—was suggested by a grouping of Booth's (group xiv.) : of course 1 person per acre is not a density associated with an urban district

in the ordinary sense of the term, but a country district cannot reach this density unless it includes a small town or portion of a town, *i.e.* unless a large proportion of its inhabitants live under urban conditions.

15.9. *Example 15.2.*—The subject of investigation is the inheritance of fertility in man.   (*Cf.* Pearson and others, ref. (323).)

Fertility in man (*i.e.* the number of children born to a given pair) is very largely influenced by the age of husband and wife at marriage (especially the latter), and by the duration of marriage.   It is desired to find whether it is also influenced by the heritable constitution of the parents, *i.e.* whether, allowance being made for the effect of such disturbing causes as age and duration of marriage, fertility is itself a heritable character.

The effect of duration of marriage may be largely eliminated by excluding all marriages which have not lasted, say, 15 years at least.   This will rather heavily reduce the number of records available, but will leave a sufficient number for discussion.   It would be desirable to eliminate the effect of late marriages in the same way by excluding all cases in which, say, husband was over 30 years of age or wife over 25 (or even less) at the time of marriage.   But, unfortunately, this is impossible ; the age of the wife—the most important factor—is only exceptionally given in peerages, family histories and similar works, from which the data must be compiled. All marriages lasting 15 years or more must therefore be included, whatever the age of the parents at marriage, and the effect of the varying age at marriage must be estimated afterwards.

15.10.   But the correlation between (1) number of children of a woman and (2) number of children of her daughter will be further affected according as we include in the record all her available daughters or only one. Suppose, *e.g.*, the number of children in the first generation is 5 (say the mother and her brothers and sisters), and the mother has three daughters with 0, 2 and 4 children respectively : are we to enter all three pairs (5, 0), (5, 2), (5, 4) in the correlation table, or only one pair ?   If the latter, which pair ?   For theoretical simplicity the second process is distinctly the better (though it still further limits the available data).   If it be adopted, some regular rule will have to be made for the selection of the daughter whose fertility shall be entered in the table, so as to avoid bias : the first daughter married for whom data are given, and who fulfils the conditions as to duration of marriage, may, for instance, be taken in every case.   (For a much more detailed discussion of the problem, and the allied problems regarding the inheritance of fertility in the horse, the student is referred to the original.)

15.11.   *Example 15.3.*– The subject for investigation is the relation between the bulk of a crop (wheat and other cereals, turnips and other root crops, hay, etc.) and the weather.   (*Cf.* Hooker, ref. (316).)

Produce statistics for the more important crops of Great Britain have been issued by the Ministry of Agriculture since 1885 : the figures are based on estimates of the yield furnished by official local estimators all over the country.   Estimates are published for separate counties and for groups of counties (divisions).   The climatic conditions vary so much over the United Kingdom that it is best to deal with a limited area, homogeneous as far as possible from the meteorological standpoint.   On the other hand, the area should not be too small ; it should be large enough to present a representative variety of soil.   The group of eastern counties, consisting of Lincoln,

Hunts, Cambridge, Norfolk, Suffolk, Essex, Bedford and Hertford, was selected as fulfilling these conditions. The group includes the county with the largest acreage of each of the ten crops investigated, with the single exception of permanent grass.

15.12. The produce of a crop is dependent on the weather of a long preceding period, and it is naturally desired to find the influence of the weather at successive stages during this period, and to determine, for each crop, which period of the year is of most critical importance as regards weather. It must be remembered, however, that the times of both sowing and harvest are themselves very largely dependent on the weather, and consequently, on an average of many years, the limits of the critical period will not be very well defined. If, therefore, we correlate the produce of the crop (X) with the characteristics of the weather (Y) during successive intervals of the year, it will be as well not to make these intervals too short. It was accordingly decided to take successive groups of 8 weeks, overlapping each other by 4 weeks, *i.e.* weeks 1–8, 5–12, etc. Correlation coefficients were thus obtained at 4-week intervals, but based on 8 weeks' weather.

15.13. It remains to be decided what characteristics of the weather are to be taken into account. The rainfall is clearly one factor of great importance, temperature is another, and these two will afford quite enough labour for a first investigation. The weekly rainfalls were averaged for eight stations within the area, and the average taken as the first characteristic of the weather. Temperatures were taken from the records of the same stations. The average temperatures, however, do not give quite the sort of information that is required : at temperatures below a certain limit (about 42° Fahr.) there is very little growth, and the growth increases in rapidity as the temperature rises above this point (within limits). It was therefore decided to utilise the figures for " accumulated temperatures above 42° Fahr.," *i.e.* the total number of day-degrees above 42° during each of the 8-weekly periods, as the second characteristic of the weather ; these " accumulated temperatures," moreover, show much larger variations than mean temperatures.

The student should refer to the original for the full discussion as to data.

The Variate-difference Correlation Method.

15.14. Problems of a somewhat special kind arise when dealing with the relations between simultaneous values of two variables which have been observed during a considerable period of time, for the more rapid movements will often exhibit a fairly close consilience, while the slower changes show no similarity. The two following examples will serve as illustrations of two methods which are generally applicable to such cases :—

Example 15.4.—Fig. 15.1 exhibits the movements of (1) the infantile mortality (deaths of infants under 1 year of age per 1000 births in the same year), (2) the general mortality (deaths at all ages per 1000 living), in England and Wales during the period 1838–1914. A very cursory inspection of the figure shows that when the infantile mortality rose from one year to the next the general mortality also rose, as a rule ; and similarly, when the infantile mortality fell, the general mortality also fell. There were, in fact, only seven or eight exceptions to this rule during the whole period under review. The correlation between the annual values of the two mortalities would nevertheless not be very high, as the general mortality
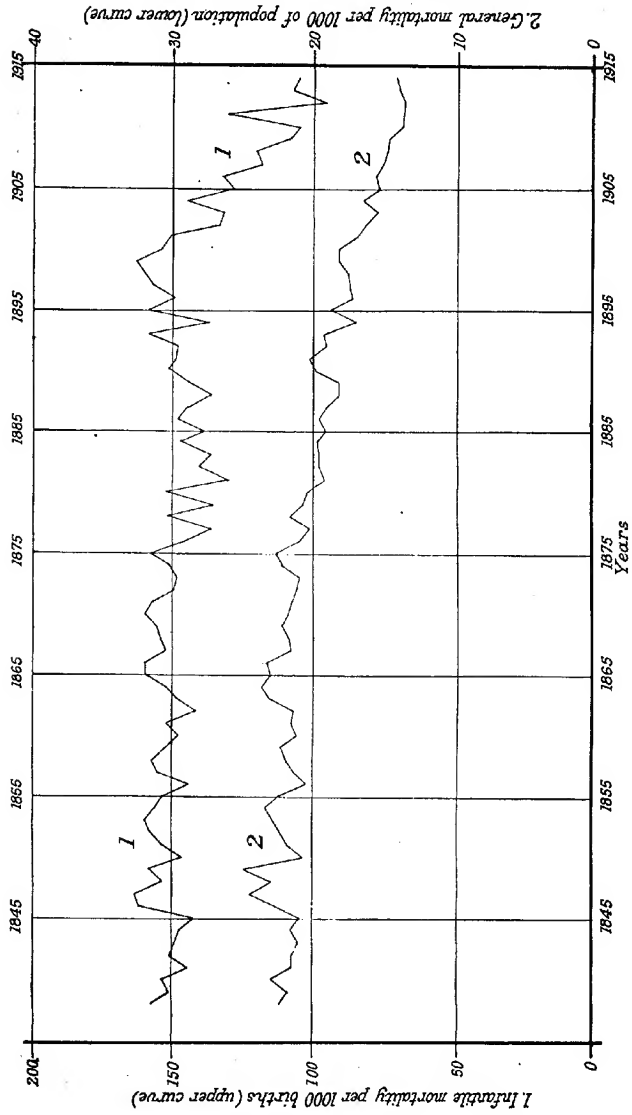
FIG. 15.1.—Infantile and General Mortality in England and Wales, 1888–1914.

has been falling more or less steadily since 1875 or thereabouts, while the infantile mortality attained almost a record value in 1898. During a long period of time the correlation between annual values may, indeed, very well vanish, for the two mortalities are affected by causes which are to a large extent different in the two cases. To exhibit, therefore, the closeness of the relation between infantile and general mortality, *for such causes as show marked changes between one year and the next,* it will be best to proceed by correlating the annual *changes,* and not the annual values. The work would be arranged in the following form (only sufficient years being given to exhibit the principle of the process), and the correlation worked out between the figures of columns 3 and 5 :—

.1.	2.	3.	4.	5.
Year.	Infantile Mortality per 1000 Births.	Increase or Decrease from Year before.	General Mortality per 1000 living.	Increase or Decrease from Year before.
1838	159	—	22·4	—
1839	151	− 8	21·8	− 0·6
1840	154	+3	22·9	+1·1
1841	145	− 9	21·6	−1·3
1842	152	+7	21·7	+0·1
1843	150	− 2	21·2	− 0·5

For the period to which the diagram refers, viz. 1838–1914, the following constants were found by this method :—

Infantile mortality, mean annual change — 0·71
       ,,   ,,   , standard deviation   10·76
General mortality, mean annual change — 0·11
       ,,   ,,   , standard deviation   1·13
Coefficient of correlation   + 0·69

This is a much higher correlation than would arise from the mere fact that the deaths of infants form part of the general mortality, and consequently there must be a high correlation between the annual changes in the mortality of those who are over and under 1 year of age, respectively. (*Cf.* Exercise 16.6, page 308.)

15.15. The procedure of the foregoing section has been called the " variate-difference correlation method." By taking first differences instead of the variate values themselves, the slower changes of the two variates with time are to some extent eliminated, and we are able to study the effect of short-term variations. To eliminate the secular changes more completely it may be desirable to proceed to second differences, *i.e.* to work out the successive differences of the differences in column 3 and column 5 before correlating. It may even be desirable to proceed to third, fourth or higher differences before correlating. The method should, however, be used with caution in such cases, particularly with short series. Correlation coefficients obtained from higher differences are not always reliable, and their interpretation becomes a matter of considerable difficulty.

15.16. *Example 15.5.*—The two curves of fig. 15.2 show (1) the marriage-rate (persons married per 1000 of the population) for England and Wales ; (2) the values of exports and imports per head of the population of the United Kingdom for every year from 1855 to 1904. Inspection of the diagram suggests a similar relation to that of the last example, the one

variable showing a rise from one year to the next when the other rises, and
a fall when the other falls. The movement of both variables is, however,
of a much more regular kind than that of mortality, resembling a series of
" waves " superposed on a steady general trend, and it is the " waves " in
the two variables—the short-period movements, not the slower trends—
which are so clearly related.

15.17. It is not difficult, moreover, to separate the short-period
oscillations, more or less approximately, from the slower movement.
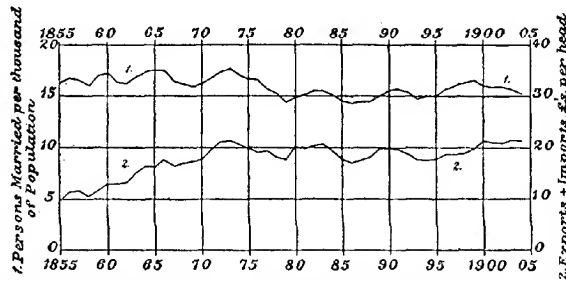
FIG. 15.2.—Marriage-rate and Foreign Trade, England and Wales, 1855-1904.

Suppose the marriage-rate for each year replaced by the average of an odd
number of years of which it is the centre, the number being as near as may
be the same as the period of the " waves "—e.g. nine years. If these short-
period averages were plotted on the diagram instead of the rates of the
individual years, we should evidently obtain a smoother curve which would
clearly exhibit the trend and be practically free from the conspicuous waves.
The excess or defect of each annual rate above or below the trend, if plotted
separately, would therefore give the " waves " apart from the slower
changes. The figures for foreign trade may be treated in the same way as
the marriage-rate, and we can accordingly work out the correlation between
the waves or rapid fluctuations, undisturbed by the movements of longer
period, however great they may be. The arithmetic may be carried out
in the form of the following table, and the correlation worked out in the
ordinary way between the figures of columns 4 and 7:—

1. Year.	2. Marriage-rate (England and Wales).	3. Nine Years' Average.	4. Differ- ence.	5. Exports+Im- ports, £'s per head (U.K.).	6. Nine Years' Average.	7. Differ- ence.
1855	16·2	—	—	9·36	—	—
1856	16·7	—	—	11·14	—	—
1857	16·5	—	—	11·85	—	—
1858	16·0	—	—	10·73	—	—
1859	17·0	16·5	+0·5	11·72	12·15	−0·43
1860	17·1	16·6	+0·5	13·03	12·94	+0·09
1861	16·3	16·7	−0·4	13·01	13·52	−0·51
1862	16·1	16·8	−0·7	13·40	14·17	−0·77
1863	16·8	16·9	−0·1	15·13	14·81	+0·32
1864	17·2	—	—	16·43	—	—
1865	17·5	—	—	16·37	—	—
1866	17·5	—	—	17·72	—	—
1867	16·5	—	—	16·47	—	—

15.18. Fig. 15.3 is drawn from the figures of columns 4 and 7, and shows very well how closely the oscillations of the marriage-rate are related to those of trade. For the period 1861–95 the correlation between the two oscillations (Hooker, ref. (314)) is 0·86. The method may obviously be extended by correlating the deviation of the marriage-rate in any one year with the deviation of the exports and imports of the year before, or two
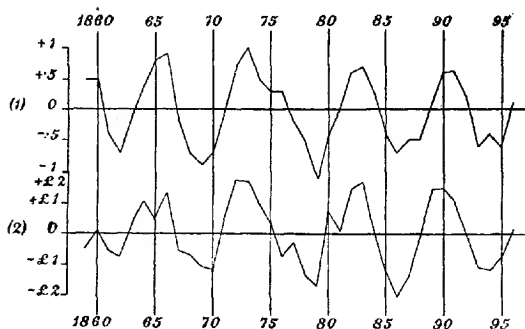
FIG. 15.3.—Fluctuations in (1) Marriage-rate and (2) Foreign Trade (Exports + Imports per head) in England and Wales: the Curves show Deviations from 9-year Means. (Data of R. H. Hooker, *Jour. Roy. Stat. Soc.*, 1901.)

years before, instead of the same year; if a sufficient number of years be taken, an estimate may be made, by interpolation, of the time-difference that would make the correlation a maximum if it were possible to obtain the figures for exports and imports for periods other than calendar years. Thus Hooker found (ref. (314)) that on an average of the years 1861–95 the correlation would be a maximum between the marriage-rate and the foreign trade of about one-third of a year earlier. The method is an extremely useful one, and is obviously applicable to any similar case. Reference may be made to ref. (335), in which several diagrams are given similar to fig. 15.3, and the nature of the relationship between the marriage-rate and such factors as trade, unemployment, etc., is discussed, it being suggested that the relation is even more complex than appears from the above.

MISCELLANEOUS THEOREMS INVOLVING THE USE OF THE CORRELATION COEFFICIENT.

Algebraical Convenience of the Correlation Coefficient.

16.1. It has already been pointed out that a statistical measure, if it is to be widely useful, should lend itself readily to algebraical treatment. The arithmetic mean and the standard deviation derive their importance largely from the fact that they fulfil this requirement better than any other averages or measures of dispersion ; and the following illustrations, while giving a number of results that are of value in one branch or another of statistical work, suffice to show that the correlation coefficient can be treated with the same facility. This might indeed be expected, seeing that the coefficient is derived, like the mean and standard deviation, by a straightforward process of summation.

The Standard Deviation of the Sum or Difference of Variables.

16.2. Let X_1, X_2 be two variables, and Z stand for their sum or difference.

Let z, x_1, x_2 denote deviations of the several variables from their arithmetic means. Then, if

$$Z = X_1 \pm X_2$$

evidently

$$z = x_1 \pm x_2$$

Squaring both sides of the equation and summing,

$$S(z^2) = S(x_1^2) + S(x_2^2) \pm 2S(x_1 x_2)$$

That is, if r be the correlation between x_1 and x_2, and σ, σ_1, σ_2 the respective standard deviations,

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 \pm 2r\sigma_1\sigma_2 \qquad . \qquad . \qquad . \quad (16.1)$$

If x_1 and x_2 are uncorrelated, we have the important special case

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 \qquad . \qquad . \qquad . \qquad . \quad (16.2)$$

The student should notice that in this case the standard deviation of the sum of corresponding values of the two variables is the same as the standard deviation of their difference.

The same process will evidently give the standard deviation of a linear function of any number of variables. *For the sum of a series of variables* $X_1, X_2, \ldots X_N$, we must have :

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_N^2 + 2r_{12}\sigma_1\sigma_2 + 2r_{13}\sigma_1\sigma_3$$
$$+ \ldots + 2r_{23}\sigma_2\sigma_3 + \ldots$$

r_{12} being the correlation between X_1 and X_2, r_{23} the correlation between X_2 and X_3, and so on.

Influence of Errors of Observation on the Standard Deviation.

16.3. The results of **16.2** may be applied to the theory of errors of observation. Let us suppose that, if *any* value of X be observed a large number of times, the arithmetic mean of the observations is approximately the true value, the arithmetic mean error being zero. Then, the arithmetic mean error being zero for all values of X, the error, say δ, is uncorrelated with X. In this case if x_1 be an observed deviation from the arithmetic mean, and x the true deviation, we have from the preceding :

$$\sigma_{x_1}^2 = \sigma_x^2 + \sigma_\delta^2 \qquad . \qquad . \qquad . \qquad . \qquad (16.3)$$

The effect of errors of observation is, consequently, to increase the standard deviation above its true value. The student should notice that the assumption made does not imply the *complete independence* of X and δ : he is quite at liberty to suppose that errors fluctuate more, for example, with large than with small values of X, as might very probably happen. In that case the contingency coefficient between X and δ would not be zero, although the correlation coefficient might still vanish as supposed.

16.4. If certain observations be repeated so that we have in every case two measures x_1 and x_2 of the same deviation x, it is possible to obtain the true standard deviation σ_x if the further assumption is legitimate that the errors δ_1 and δ_2 are uncorrelated with each other. On this assumption

$$S(x_1 x_2) = S(x + \delta_1)(x + \delta_2)$$
$$= S(x^2)$$

and accordingly

$$\sigma_x{}^2 = \frac{S(x_1 x_2)}{N} \qquad . \qquad . \qquad . \qquad . \qquad (16.4)$$

(This formula is part of Spearman's formula for the correction of the correlation coefficient; *cf.* **16.6.**)

Influence of Errors of Observation on the Correlation Coefficient.

16.5. Let x_1, y_1 be the observed deviations from the arithmetic means, x, y the true deviations, and δ, ϵ the errors of observation. Of the four quantities x, y, δ, ϵ we will suppose x and y alone to be correlated. On this assumption

$$S(x_1 y_1) = S(xy) \qquad . \qquad . \qquad . \qquad . \qquad (16.5)$$

It follows at once that

$$\frac{r_{xy}}{r_{x_1 y_1}} = \frac{\sigma_{x_1} \sigma_{y_1}}{\sigma_x \sigma_y}$$

and consequently the observed correlation is less than the true correlation. This difference, it should be noticed, no mere increase in the number of observations can in any way lessen.

Spearman's Theorems.

16.6. If, however, the observations of both x and y be repeated, as assumed in **16.4**, so that we have two measures x_1 and x_2, y_1 and y_2 of every value of x and y, the true value of the correlation can be obtained by the use of equations (16.4) and (16.5), on assumptions similar to those made above.   For we have:

$$r^2_{xy} = \frac{S(x_1 y_1) S(x_2 y_2)}{S(x_1 x_2) S(y_1 y_3)} = \frac{S(x_1 y_2) S(x_2 y_1)}{S(x_1 x_2) S(y_1 y_2)}$$

$$= \frac{r_{x_1 y_1} r_{x_2 y_2}}{r_{x_1 x_2} r_{y_1 y_2}} = \frac{r_{x_1 y_2} r_{x_2 y_1}}{r_{x_1 x_2} r_{y_1 y_2}} \qquad . \qquad . \qquad . \qquad (16.6)$$

Or, if we use all the four possible correlations between observed values of x and observed values of y,

$$r^4_{xy} = \frac{r_{x_1 y_1} r_{x_2 y_2} r_{x_1 y_2} r_{x_2 y_1}}{\left(r_{x_1 x_2} r_{y_1 y_2} \right)^2} \qquad . \qquad . \qquad . \qquad (16.7)$$

Equation (16.7) is the original form in which Spearman gave his correction formula (refs. (339) and (340)).   It will be seen to imply the assumption that, of the six quantities x, y, δ_1, δ_2, ϵ_1, ϵ_2, only x and y are correlated.   The correction given by the second part of equation (16.6), also suggested by Spearman, seems, on the whole, to be safer, for it eliminates the assumption that the errors in x and in y, in the same series of observations, are uncorrelated.   An insufficient though partial test of the correctness of the assumptions may be made by correlating $x_1 - x_2$ with $y_1 - y_2$:   this correlation should vanish.   Evidently, however, it may vanish from symmetry without thereby implying that all the correlations of the errors are zero.

Mean and Standard Deviation of an Index.

16.7. The means and standard deviations of non-linear functions of two or more variables can in general only be expressed in terms of the means and standard deviations of the original variables to a first approximation, on the assumption that deviations are small compared with the mean values of the variables.   Thus, let it be required *to find the mean and standard deviation of a ratio or index* $Z = X_1 / X_2$, in terms of the constants for X_1 and X_2.   Let I be the mean of Z, M_1 and M_2 the means of X_1 and X_2.   Then,

$$I = \frac{1}{N} S\left(\frac{X_1}{X_2}\right) = \frac{1}{N} \frac{M_1}{M_2} S\left(1 + \frac{x_1}{M_1}\right)\left(1 + \frac{x_2}{M_2}\right)^{-1}$$

Expand the second bracket by the binomial theorem, assuming that x_2/M_2 is so small that powers higher than the second can be neglected. Then, to this approximation,

$$I = \frac{1}{N} \frac{M_1}{M_2}\left[N - \frac{1}{M_1 M_2} S(x_1 x_2) + \frac{1}{M_2^2} S(x_2^{\,2}) \right]$$

That is, if r be the correlation between x_1 and x_2, and if $v_1 = \sigma_1/M_1$, $v_2 = \sigma_2/M_2$,

$$I = \frac{M_1}{M_2}(1 - r v_1 v_2 + v_2^{\,2}) \qquad . \qquad . \qquad . \qquad (16.8)$$

If s be the standard deviation of Z, we have :

$$s^2 + I^2 = \frac{1}{N}S\left(\frac{X_1}{X_2}\right)^2$$

$$= \frac{1}{N}\frac{M_1^2}{M_2^2}S\left(1 + \frac{x_1}{M_1}\right)^2\left(1 + \frac{x_2}{M_2}\right)^{-2}$$

Expanding the second bracket again by the binomial theorem, and neglecting terms of all orders above the second :

$$s^2 + I^2 = \frac{1}{N}\frac{M_1^2}{M_2^2}S\left(1 + \frac{x_1}{M_1}\right)^2\left(1 - 2\frac{x_2}{M_2} + 3\frac{x_2^2}{M_2^2}\right)$$

$$= \frac{M_1^2}{M_2^2}(1 + v_1^2 - 4rv_1v_2 + 3v_2^2)$$

or from (16.8) :

$$s^2 = \frac{M_1^2}{M_2^2}(v_1^2 - 2rv_1v_2 + v_2^2) \quad . \quad\quad . \quad\quad . \quad\quad . \quad (16.9)$$

Correlation between Indices.

16.8.   The following problem affords a further illustration of the use of the same method.   *Required to find approximately the correlation between two ratios $Z_1 = X_1/X_3$, $Z_2 = X_2/X_3$, X_1, X_2 and X_3 being uncorrelated.*

Let the means of the two ratios or indices be I_1, I_2, and the standard deviations s_1, s_2; these are given approximately by (16.8) and (16.9) of the last section.   The required correlation ρ will be given by

$$N\rho s_1 s_2 = S\left(\frac{X_1}{X_3} - I_1\right)\left(\frac{X_2}{X_3} - I_2\right)$$

$$= S\left(\frac{X_1 X_2}{X_3^2}\right) - NI_1 I_2$$

$$= \frac{M_1 M_2}{M_3^2}S\left(1 + \frac{x_1}{M_1}\right)\left(1 + \frac{x_2}{M_2}\right)\left(1 + \frac{x_3}{M_3}\right)^{-2} - NI_1 I_2$$

Neglecting terms of higher order than the second as before and remembering that all correlations are zero, we have :

$$\rho s_1 s_2 = \frac{M_1 M_2}{M_3^2}(1 + 3v_3^2) - I_1 I_2$$

$$= \frac{M_1 M_2}{M_3^2}v_3^2$$

where, in the last step, a term of the order v_3^4 has again been neglected. Substituting from (16.9) for s_1 and s_2, we have finally :

$$\rho = \frac{v_3^2}{\sqrt{(v_1^2 + v_3^2)(v_2^2 + v_3^2)}} \quad . \quad\quad . \quad\quad . \quad (16.10)$$

This value of ρ is obviously positive, being equal to 0·5 if $v_1 = v_2 = v_3$; and hence even if X_1 and X_2 are independent, the indices formed by taking their ratios to a common denominator X_3 will be correlated.   The value of ρ was termed by Karl Pearson the " spurious correlation." Thus, if

measurements be taken, say, on three bones of the human skeleton, and the measurements grouped in threes absolutely at random, there will, nevertheless, be a positive correlation, probably approaching 0·5, between the indices formed by the ratios of two of the measurements to the third. To give another illustration, if two individuals both observe the same series of magnitudes quite independently, there may be little, if any, correlation between their absolute errors. But if the errors be expressed as percentages of the magnitude observed, there may be considerable correlation. It does not follow of necessity that the correlations between indices or ratios are misleading. If the indices are uncorrelated, there will be a similar " spurious " correlation between the absolute measurements $Z_1 X_3 = X_1$ and $Z_2 X_3 = X_2$, and the answer to the question whether the correlation between indices or that between absolute measures is misleading depends on the further question whether the indices or the absolute measures are the quantities directly determined by the causes under investigation (*cf.* ref. (346)).

The case considered, where X_1, X_2, X_3 are uncorrelated, is only a special one ; for the general discussion *cf.* ref. (345). For an interesting study of actual illustrations *cf.* ref. (343).

Correlation due to Heterogeneity of Material.

16.9. The following theorem offers some analogy with the theorem of **4.12** for attributes: *If X and Y are uncorrelated in each of two records, they will nevertheless exhibit some correlation when the two records are mingled, unless the mean value of X in the second record is identical with that in the first record, or the mean value of Y in the second record is identical with that in the first record, or both.*

This follows almost at once, for if M_1, M_2 are the mean values of X in the two records, K_1, K_2 the mean values of Y, N_1, N_2 the numbers of observations, and M, K the means when the two records are mingled, the product-sum of deviations about M, K is

$$N_1(M_1 - M)(K_1 - K) + N_2(M_2 - M)(K_2 - K)$$

Evidently the first term can only be zero if $M = M_1$ or $K = K_1$. But the first condition gives

$$\frac{N_1 M_1 + N_2 M_2}{N_1 + N_2} = M_1$$

that is,

$$M_1 = M_2$$

Similarly, the second condition gives $K_1 = K_2$. Both the first and second terms can, therefore, only vanish if $M_1 = M_2$ or $K_1 = K_2$. Correlation may accordingly be created by the mingling of two records in which X and Y vary round different means. (For a more general form of the theorem *cf.* ref. (323).)

Reduction of Correlation due to Mingling of Uncorrelated with Correlated Pairs.

16.10. Suppose that n_1 observations of x and y give a correlation coefficient

$$r_1 = \frac{S(xy)}{n_1 \sigma_x \sigma_y}$$

Now, let n_2 pairs be added to the material, the means and standard deviations of x and y being the same as in the first series of observations, but the correlation zero. The value of $S(xy)$ will then be unaltered, and we will have:

$$r_2 = \frac{S(xy)}{(n_1 + n_2)\sigma_x\sigma_y}$$

Whence

$$\frac{r_2}{r_1} = \frac{n_1}{n_1 + n_2} \qquad . \qquad . \qquad . \qquad . \qquad (16.11)$$

Suppose, for example, that a number of bones of the human skeleton have been disinterred during some excavations, and a correlation r_2 is observed between pairs of bones presumed to come from the same skeleton, this correlation being rather lower than might have been expected, and subject to some uncertainty owing to doubts as to the allocation of certain bones. If r_1 is the value that would be expected from other records, the difference might be accounted for on the hypothesis that, in a proportion $(r_1 - r_2)/r_1$ of all the pairs, the bones do not really belong to the same skeleton, and have been virtually paired at random.

The Weighted Mean.

16.11. The arithmetic mean M of a series of values of a variable X was defined as the quotient of the sum of those values by their number N, or

$$M = S(X)/N$$

If, on the other hand, we multiply each individual observed value of X by some numerical coefficient or *weight* W, the quotient of the sum of such products by the sum of the weights is defined as a *weighted mean* of X, and may be denoted by M'; so that

$$M' = S(WX)/S(W)$$

The distinction between " weighted " and " unweighted " means is, it should be noted, very often formal rather than essential, for the " weights " may be regarded as actual, estimated or virtual frequencies. The weighted mean then becomes simply an arithmetic mean, in which some new quantity is regarded as the unit. Thus, if we are given the means $M_1, M_2, M_3, \ldots M_r$ of r series of observations, but do not know the number of observations in every series, we may form a general average by taking the arithmetic mean of all the means, viz. $S(M)/r$, treating the series as the unit. But if we know the number of observations in every series it will be better to form the *weighted mean* $S(NM)/S(N)$, weighting each mean in proportion to the number of observations in the series on which it is based. The second form of average would be quite correctly spoken of as a weighted mean of the means of the several series : at the same time, it is simply the arithmetic mean of all the series pooled together, *i.e.* the arithmetic mean obtained by treating the observation and not the series as the unit.

16.12. To give an arithmetical illustration, if a commodity is sold at different prices in different markets, it will be better to form an average price, not by taking the arithmetic mean of the several market prices,

treating the market as the unit, but by weighting each price in proportion to the quantity sold at that price, if known, *i.e.* treating the unit of quantity as the unit of frequency. Thus, if wheat has been sold in market A at an average price of 29s. 1d. per quarter, in market B at an average price of 27s. 7d. and in market C at an average price of 28s. 4d., we may, if no statement is made as to the quantities sold at these prices (as very often happens in the case of statements as to market prices), take the arithmetic mean (28s. 4d.) as the general average. But if we know that 23,930 qrs. were sold at A, only 26 qrs. at B and 3,933 qrs. at C, it will be better to take the *weighted mean*

$$\frac{(29s. \ 1d. \times 23,930) + (27s. \ 7d. \times 26) + (28s. \ 4d. \times 3,933)}{27,889} = 29s.$$

to the nearest penny. This is appreciably higher than the arithmetic mean price, which is lowered by the undue importance attached to the small markets B and C.

16.13. In the case of index-numbers for exhibiting the changes in average prices from year to year (*cf.* 7.34), it may make a sensible difference whether we take the simple arithmetic mean of the index-numbers for different commodities in any one year as representing the price-level in that year, or *weight* the index-numbers for the several commodities according to their importance from some point of view; and much has been written as to the weights to be chosen. If, for example, our standpoint be that of some average consumer, we may take as the *weight* for each commodity the sum which he spends on that commodity in an average year, so that the frequency of each commodity is taken as the number of shillings or pounds spent thereon instead of simply as unity.

16.14. Rates or ratios like the birth-, death- or marriage-rates of a country may be regarded as weighted means. For, treating the rate for simplicity as a fraction, and not as a rate per 1000 of the population,

$$\text{Birth-rate of whole country} = \frac{\text{Total births}}{\text{Total population}}$$

$$= \frac{S(\text{Birth-rate in each district} \times \text{population in that district})}{S(\text{Population of each district})}$$

i.e. the rate for the whole country is the mean of the rates in the different districts, weighting each in proportion to its population. We use the weighted and unweighted means of such rates as illustrations in **16.16** below.

16.15. It is evident that any weighted mean will in general differ from the unweighted mean of the same quantities, and it is required to find an expression for this difference. If r be the correlation between weights and variables, σ_w and σ_x the standard deviations and $\bar{w}$ the mean weight, we have at once

$$S(WX) = N(M\bar{w} + r\sigma_w\sigma_x)$$

whence

$$M' = M + r\sigma_x\frac{\sigma_w}{\bar{w}} \qquad . \qquad . \qquad . \qquad (16.12)$$

That is to say, if the weights and variables are positively correlated, the weighted mean is the greater ; if negatively, the less. In some cases r is very small, and then weighting makes little difference, but in others the difference is large and important, r having a sensible value and $\sigma_x\sigma_w/\bar{w}$ a large value.

16.16. The difference between weighted and unweighted means of death-rates, birth-rates or other rates on the population in different districts is, for instance, nearly always of importance. Thus we have the following figures for rates of pauperism (*Jour. Roy. Stat. Soc.*, vol. 59, 1896, p. 349) :—

January 1.	Percentages of the Population in receipt of Relief.	
	Arithmetic Mean of Rates in different Districts.	England and Wales as a whole.
1850	6·51	5·80
1860	5·20	4·26
1870	5·45	4·77
1881	3·68	3·12
1891	3·29	2·69

In this case the weighted mean is markedly the less, and the correlation between the population of a district and its pauperism must therefore be negative, the larger (on the whole urban) districts having the lower percentage in receipt of relief. On the other hand, for the decade 1881–90 the average birth-rate for England and Wales was 32·34 per thousand, the arithmetic mean of the rates for the different districts 30·34 only. The weighted mean was therefore the greater, the birth-rate being higher in the more populous (urban) districts, in which there is a greater proportion of young married persons.

For the year 1891 the average population of a poor law district was found to be roughly 45,900 and the standard deviation σ_w 56,400 (populations ranging from under 2000 to over half a million). The standard deviation σ_x of the percentages of the population in receipt of relief was 1·24. We have therefore, for the correlation between pauperism and population,

$$r = -\frac{3\cdot29 - 2\cdot69}{1\cdot24} \times \frac{459}{564}$$
$$= -0\cdot39$$

For the birth-rate, on the other hand, assuming that $\sigma_w/\bar{w}$ is approximately the same for the decade 1881–90 as in 1891, and neglecting the fact that in a few instances Registration Districts differ from Poor-law Unions, we have, σ_x being 4·08,

$$r = \frac{32\cdot34 - 30\cdot34}{4\cdot08} \times \frac{459}{564}$$
$$= +0\cdot40$$

The closeness of the numerical values of r in the two cases is, of course, accidental.

16.17. The principle of weighting finds one very important application in the treatment of such rates as death-rates, which are largely affected by the age and sex composition of the population. Neglecting, for simplicity, the question of sex, suppose the numbers of deaths are noted in a certain district for, say, the age-groups 0–, 10–, 20–, etc., in which the fractions of the whole population are p_0, p_1, p_2, etc., where $S(p) = 1$. Let the death-rates for the corresponding age-groups be d_0, d_1, d_2, etc. Then the ordinary or *crude* death-rate for the district is

$$D = S(dp) \qquad . \qquad . \qquad . \qquad . \qquad (16.13)$$

For some other district taken as a basis of comparison, perhaps the country as a whole, the death-rates and fractions of the population in the several age-groups may be δ_1, δ_2, δ_3, . . ., π_1, π_2, π_3, . . ., and the crude death-rate

$$\Delta = S(\delta\pi) \qquad . \qquad . \qquad . \qquad (16.14)$$

Now, D and Δ differ either because the d's and δ's differ or because the p's and π's differ, or both. It may happen that really both districts are about equally healthy, and the death-rates approximately the same for all age-classes, but, owing to a difference of *weighting*, the first average may be markedly higher than the second, or *vice versa*. If the first district be a rural district and the second urban, for instance, there will be a larger proportion of the old in the former, and it may possibly have a higher crude death-rate than the second, in spite of lower death-rates in every class. The comparison of crude death-rates is therefore liable to lead to erroneous conclusions. The difficulty may be got over by averaging the age-class death-rates in the district not with the weights p_1, p_2, p_3, . . . given by its own population, but with the weights π_1, π_2, π_3, . . . given by the population of the standard district. The *standardised death-rate* for the district will then be

$$D' = S(d\pi) \qquad . \qquad . \qquad . \qquad (16.15)$$

and D' and Δ will be comparable as regards age-distribution. There is obviously no difficulty in taking sex into account as well as age if necessary. The death-rates must be noted for each sex separately in every age-class and averaged with a system of weights based on the standard population. The method is also of importance for comparing death-rates in different classes of the population, *e.g.* those engaged in given occupations, as well as in different districts, and is used for both these purposes in the publications of the Registrar-General for England and Wales.

16.18. Difficulty may arise in practical cases from the fact that the death-rates d_1, d_2, d_3, . . . are not known for the districts or classes which it is desired to compare with the standard population, but only the crude rates D and the fractional populations of the age-classes p_1, p_2, p_3, . . . The difficulty may be partially obviated (*cf.* **4.16** and Example 4.3, pp. 58–60) by forming what is termed an *index* death-rate Δ' for the class or district, Δ' being given by

$$\Delta' = S(\delta p) \qquad . \qquad . \qquad . \qquad (16.16)$$

i.e. the rates of the standard population averaged with the weights of

the district population. It is the crude death-rate that there would be in the district if the rate in every age-class were the same as in the standard population. An approximate standardised death-rate for the district or class is then given by

$$D'' = D \times \frac{\Delta}{\Delta'} \qquad . \qquad . \qquad . \qquad . \qquad (16.17)$$

D'' is not necessarily, nor generally, the same as D'. It can only be the same if

$$\frac{S(d\pi)}{S(dp)} = \frac{S(\delta\pi)}{S(\delta p)}$$

This will hold good if, *e.g.*, the death-rates in the standard population and the district stand to one another in the same ratio in all age-classes, *i.e.* $\delta_1/d_1 = \delta_2/d_2 = \delta_3/d_3 = $ etc. This method of standardisation was used in the Annual Summaries of the Registrar-General for England and Wales.

16.19. Both methods of standardisation—that of **16.17** and that of **16.18**—are of great importance. They are obviously applicable to other rates besides death-rates, *e.g.* birth-rates. Further, they may readily be extended into quite different fields. Thus it has been suggested that standardised *average heights* or standardised *average weights* of the children in different schools might be obtained on the basis of a standard school population of given age and sex composition, or indeed of given composition as regards hair- and eye-colour as well.

16.20. In **16.11–16.16** we have dealt only with the theory of the weighted arithmetic mean, but it should be noted that any form of average can be weighted. Thus a weighted median can be formed by finding the value of the variable such that the sum of the weights of lesser values is equal to the sum of the weights of greater values. A weighted mode could be formed by finding the value of the variable for which the sum of the weights was greatest, allowing for the smoothing of casual fluctuations. Similarly, a weighted geometric mean could be calculated by weighting the logarithms of every value of the variable before taking the arithmetic mean, *i.e.*

$$\log G_w = \frac{S(W \log X)}{S(W)}$$

SUMMARY.

1. The standard deviation of the sum of variables $X_1, X_2, \ldots X_N$ is given by

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_N^2 + 2r_{12}\sigma_1\sigma_2 + 2r_{13}\sigma_1\sigma_3 + \ldots + 2r_{23}\sigma_2\sigma_3 + \ldots$$

2. In particular, the variance of the sum of N uncorrelated variates is the sum of their variances.

3. If X_1, X_2 and X_3 are uncorrelated, the indices $\frac{X_1}{X_3}, \frac{X_2}{X_3}$ will nevertheless be correlated in general.

4. If X and Y are uncorrelated in each of two separate records, they will be correlated in the sum of the two records, unless either the means of X or the means of Y, or both, are the same in the two records.

5. If correlated and uncorrelated material is mingled, the correlation in the total is lower than that in the correlated portion.

6. An arithmetic mean is weighted when, in the calculation of $\frac{1}{N}S(X)$, each value of the variate is multiplied by a weight W.

7. The weighted arithmetic mean is greater or less than the unweighted mean according as the weights and variables are positively or negatively correlated.

EXERCISES.

16.1. (Data from the Decennial Supplements to the Annual Reports of the Registrar-General for England and Wales.) The following particulars are found for 36 small registration districts in which the number of births in a decade ranged between 1500 and 2500:—

Decade.	Proportion of Male Births per 1000 of all Births.	
	Mean.	Standard deviation.
1881–1890 . .	508·1	12·80
1891–1900 . .	508·4	10·37
Both decades	508·25	11·65

It is believed, however, that a great part of the observed standard deviation is due to mere "fluctuations of sampling" of no real significance.

Given that the correlation between the proportions of male births in a district in the two decades is $+0.36$, estimate (1) the true standard deviation freed from such fluctuations of sampling; (2) the standard deviation of fluctuations of sampling, *i.e.* of the errors produced by such fluctuations in the observed proportions of male births.

16.2. (Data from Pearson, ref. (345).) The coefficients of variation for breadth, height and length of certain skulls are 3·89, 3·50 and 3·24 per cent. respectively. Find the "spurious correlation" between the breadth/length and height/length indices, absolute measures being combined at random so that they are uncorrelated.

16.3. (Data from Boas, communicated to Pearson; *cf.* Fawcett and Pearson, *Proc. Roy. Soc.*, vol. 62, p. 413.) From short series of measurements on American Indians, the mean coefficient of correlation found between father and son, and father and daughter, for cephalic index, is 0·14; between mother and son, and mother and daughter, 0·33. Assuming these coefficients should be the same if it were not for the looseness of family relations, find the proportion of children not due to the reputed father.

16.4. Find the correlation between $X_1 + X_2$ and $X_2 + X_3$, X_1, X_2 and X_3 being uncorrelated.

16.5. Find the correlation between X_1 and $aX_1 + bX_2$, X_1 and X_2 being uncorrelated.

16.6. (Referring to Example 15.4, p. 292.)   Use the answer to Exercise 16.5 to estimate, very roughly, the correlation that would be found between annual movements in infantile and general mortality if the mortality of those under and over 1 year of age were uncorrelated.   Note that—

$$\left.\begin{array}{l}\text{General mortality per}\\\quad\text{1000 of population}\end{array}\right\} = \text{Infantile mortality per 1000 births} \times \dfrac{\text{Births}}{\text{Population}}$$

$$+ \text{Deaths over one year per 1000 of population}$$

and treat the ratio of births to population as if it were constant at a rough average value, say 0·032. The standard deviation of annual movements in infantile mortality is (*loc. cit.*) 10·76, and that of annual movements in mortality other than infantile may be taken as sensibly the same as that of general mortality, or, say, 1·13 units.

16.7. If the relation

$$ax_1 + bx_2 + cx_3 = 0$$

holds for all values of x_1, x_2 and x_3 (which are, in our usual notation, deviations from the respective arithmetic means), find the correlations between x_1, x_2 and x_3 in terms of their standard deviations and the values of a, b and c.

16.8. What is the effect on a weighted mean of errors in the weights of the quantities weighted, such errors being uncorrelated with one another, with the weights or with the variables: (1) if the arithmetic mean values of the errors are zero, (2) if the arithmetic mean values of the errors are not zero?

16.9. The following are the variances of the rainfall (1) for January to March, (2) for April to December, (3) for the whole year, at Greenwich in the eighty years 1841–1920, the unit being a millimetre : -

January–March	.	.	.	.	$\sigma_1{}^2 = 1{,}521$
April–December	.	.	.	.	$\sigma_2{}^2 = 8{,}968$
Whole year	.	.	.	.	$\sigma^2 = 10{,}754$

Find the correlation between the rainfall in January–March and April–December.

SIMPLE CURVE FITTING.

The Problem.

17.1. In this chapter we turn aside somewhat from the line of development of previous chapters in order to study a subject of considerable theoretical and practical importance—the representation of relationship between two variables by simple algebraic expressions. Our work on correlation has already led us to fit regression lines and planes to the means of arrays. We now attack a rather more general problem. An illustration will make clear the type of inquiry involved.

Table 17.1 shows the estimated distance and velocities of recession of certain nebulæ in the outlying parts of the visible universe.

TABLE 17.1.—*Estimated Distance and Velocities of Recession of 10 Extra-galactic Nebulæ* (Edwin Hubble and Milton L. Humason, "The Velocity-distance Relation among Extra-galactic Nebulæ," *Contributions from Mount Wilson Observatory*, Carnegie Institute of Washington, No. 427; *Astrophysical Journal*, vol. 74, 1931, pp. 43–80).

Constellation in which the Nebula is situated.	Mean Velocity (kilometres per second).	Distance (millions of parsecs).
Isolated Nebula II .	630	1·20
Virgo . . .	890	1·82
Isolated Nebula I .	2,350	3·31
Pegasus . .	3,810	7·24
Pisces . . .	4,630	6·92
Cancer . .	4,820	9·12
Perseus . .	5,230	10·97
Coma . .	7,500	14·45
Ursa Major . .	11,800	22·91
Leo . . .	19,600	36·31

A little inspection of the table will show that there appears to be some relation between distance and velocity—the greater the one, the greater the other, with only one exception. A diagram makes the relation clearer still. In fig. 17.1 we have taken the two variables velocity and distance as rectangular co-ordinates y and x, and have marked for each nebula a point whose co-ordinates are the distance and velocity of that nebula. The ten points so obtained evidently lie very approximately on a straight line or, to express the same fact algebraically, the ten values of the variables are closely represented by an equation of the form

$$y = a_0 + a_1 x \qquad . \qquad . \qquad . \qquad . \quad (17.1)$$

17.2. No straight line, however, passes exactly through all the points, although a great many lines may be drawn which nearly do so. The question then arises, is there a straight line which fits the points better than all others, and if so, which is it ? Or, in other language, what values of a_0 and a_1 in equation (17.1) must we take to get the best representation of the linear relationship between the two variables ? And, as a further question, can we devise a measure of the closeness of the fit of the various lines which can be drawn ?
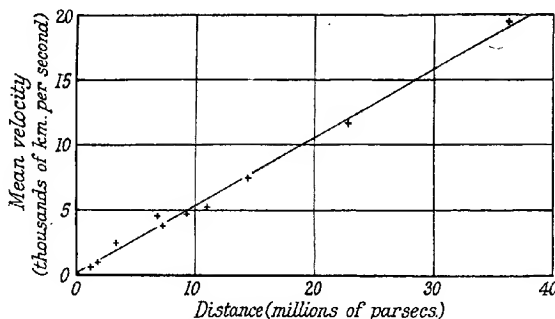
FIG. 17.1.—Relationship between Distance and Velocity of Recession in Certain Extra-galactic Nebulæ. (Table 17.1.)

17.3. In the foregoing illustration it is clear from the data or from the diagram that a linear relationship between the variables gives a very close picture of the truth. In other cases the points of the diagram will lie more or less on a curve, and no straight line will give a satisfactory representation. We should then wish to investigate whether the dependence of y on x may be suitably represented by the more general equation

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_p x^p . \qquad . \qquad . \quad (17.2)$$

which, in the diagram, corresponds to a curve of the type known as parabolic. The number p indicates the degree of the parabola, and we speak of quadratic, cubic, quartic parabolas, meaning curves of type (17.2) with $p = 2$, 3, 4, respectively.

17.4. Our general problem may, then, be stated as follows : Given n pairs of values of two variables, $X_1 Y_1$, $X_2 Y_2$, . . . $X_n Y_n$, to express the values of one of them as nearly as may be in terms of the other by an equation of the form (17.2) ; and to measure the closeness of the approximation of the values of y given by the equation to the actual values. In geometrical language, given n points in a plane, to fit to them a curve of the parabolic type (17.2) and to measure the closeness of fit.

17.5. The representation of data in this way may serve several purposes. In the first place, it may present the relationship between the two variables in a useful summary form. Secondly, it may be used to interpolate, *i.e.* to estimate the values of one variable which would correspond to specified values of the other. In fig. 17.1, for example, the straight line which has been drawn in, and whose equation is obtained below, tells us what we might expect to be the velocity of a nebula whose

distance is, say, 20 million parsecs, on the assumption that the linear relation holds good for nebulæ in general.

17.6. Again, the representation may also be very suggestive to the theorist. The linear form of the relationship between the variables of Table 17.1 means more than a convenient summary of the facts, and has inspired a great deal of research into the nature of the physical universe. In such cases, the derived equation is regarded as the expression of a law of nature, and the deviations of the observed values from those given by it are interpreted as fluctuations arising from experimental error or secondary perturbations. This standpoint is common in physics, in which data often lie very closely about a smooth curve.

The Method of Least Squares.

17.7. Let us suppose that we have n pairs of values $X_1 Y_1, \ldots X_n Y_n$, and that we wish to represent them by an equation of the type (17.2). Our problem is, having fixed the value of p, to determine the constants $a_0, a_1, \ldots a_p$ in terms of the observed values X, Y, so as to get the best possible fit.

The expression " best possible fit " may be defined in more than one way, and consequently there is no unique method of determining the constants. Several methods have been proposed, and our choice between them is determined mainly by convenience. One way, which is suggested by the geometrical representation, is to choose the curve of equation (17.2) so that the sum of the distances (taken as positive) of the points from it is a minimum, the sum of the distances being regarded as a measure of goodness of fit, and the "best" fit being given by the curve of specified degree for which that sum is least. But this method, whatever its theoretical attractions, suffers from the disadvantage that it is difficult to apply in practice except for the straight line.

An alternative method, which is in almost universal use at the present time, is that known as the **Method of Least Squares,** and we proceed to discuss it at length. We have already used it to find regression lines **(11.20 and 14.4).**

17.8. If we substitute the value X_r in equation (17.2) we get a quantity y_r, given by

$$y_r = a_0 + a_1 X_r + a_2 X_r^2 + \ldots + a_p X_r^p \qquad . \qquad (17.3)$$

This is not in general the same as Y_r, and we therefore define the residual ξ_r as

$$\xi_r = Y_r - y_r = Y_r - a_0 - a_1 X_r - \ldots - a_p X_r^p \qquad . \qquad (17.4)$$

There will be n residuals, one for each pair X, Y, and they are all zero if, and only if, the curve is a perfect fit. We then take the sum of the squares of residuals :

$$^* \; U = S(\xi_r^2) = S(Y_r - a_0 - a_1 X_r - \ldots - a_p X_r^p)^2 \qquad . \qquad (17.5)$$

If U is zero, each residual must be zero, and the data are represented perfectly by the equation. Except in this case, U is positive. The further the points lie from the curve of equation (17.2), the greater U will be. U therefore provides one measure of the closeness of fit. From this standpoint, the best fit will be that for which U is least.

The Method of Least Squares adopts this criterion, and states that *the constants* a *shall be determined so that* U *is a minimum*.

17.9. The reason for taking the sum of *squares* of residuals, rather than the sum of residuals simply, is akin to that which led us to prefer the standard deviation to the mean deviation as a measure of dispersion (Chap. 8), namely, that the former is more convenient in theory and leads to equations which are easier to handle in practice.

17.10. It was formerly the custom, and is so still in works on the theory of observations, to derive the method of least squares from certain theoretical considerations, the assumed normality of the distribution of errors of observation being one such. It is, however, more than doubtful whether the conditions for the theoretical validity of the method are realised in statistical practice, and the student would do well to regard the method as recommended chiefly by its comparative simplicity and by the fact that it has stood the test of experience.

17.11. Consider now the quantity U, given by equation (17.5). $a_0, a_1, \ldots a_p$ are to be chosen so that this is a minimum, say U_0. Let us imagine this done.

If, now, we substitute in equation (17.5) $a_0 + \epsilon_0$ for a_0, $a_1 + \epsilon_1$ for a_1, $a_2 + \epsilon_2$ for a_2, and so on, we shall get a quantity U_1 given by

$$U_1 = S\{Y - (a_0 + \epsilon_0) - (a_1 + \epsilon_1)X - \ldots - (a_p + \epsilon_p)X^p\}^2$$

and U_1 is greater than U_0 for all values of $\epsilon_0, \epsilon_1, \ldots \epsilon_p$.

Now,

$$U_1 = S\{(Y - a_0 - a_1X - \ldots - a_pX^p) - (\epsilon_0 + \epsilon_1X + \ldots + \epsilon_pX^p)\}^2$$
$$= S(Y - a_0 - a_1X - \ldots - a_pX^p)^2$$
$$\quad - 2S(Y - a_0 - a_1X - \ldots - a_pX^p)(\epsilon_0 + \epsilon_1X + \ldots + \epsilon_pX^p)$$
$$\quad + S(\epsilon_0 + \epsilon_1X + \ldots + \epsilon_pX^p)^2$$

The first of these terms is equal to U_0. Hence, if $U_1 \geqslant U_0$, we must have

$$- 2S(Y - a_0 - a_1X - \ldots - a_pX^p)(\epsilon_0 + \epsilon_1X + \ldots + \epsilon_pX^p)$$
$$+ S(\epsilon_0 + \epsilon_1X + \ldots + \epsilon_pX^p)^2 \geqslant 0 \ . \qquad \qquad (17.6)$$

This is to be true for all values of $\epsilon_0 \ldots \epsilon_p$. Let us then take these quantities to be very small. The second term in equation (17.6), depending as it does on the squares of the ϵ's, will be small compared with the first, and may be neglected. (17.6) will then be true only if the first term vanishes, for otherwise the ϵ's could be so chosen in sign as to make the first term negative.

Hence,

$$S(Y - a_0 - a_1X - \ldots - a_pX^p)(\epsilon_0 + \epsilon_1X + \ldots + \epsilon_pX^p) = 0 \qquad (17.7)$$

This is true for *all* small values of the ϵ's. Hence the coefficients of $\epsilon_1, \ldots \epsilon_p$ all vanish, *i.e.* we have:

$$
\left.
\begin{aligned}
S(Y) \quad - a_0 n \quad - a_1 S(X) \quad - \ldots - a_p S(X^p) &= 0 \\
S(YX) \quad - a_0 S(X) - a_1 S(X^2) \quad - \ldots - a_p S(X^{p+1}) &= 0 \\
S(YX^2) - a_0 S(X^2) - a_1 S(X^3) \quad - \ldots - a_p S(X^{p+2}) &= 0 \\
\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
S(YX^p) - a_0 S(X^p) - a_1 S(X^{p+1}) - \ldots - a_p S(X^{2p}) &= 0
\end{aligned}
\right\} \quad (17.8)
$$

The equations (17.8) give us $p+1$ equations in the $(p+1)$ unknowns $a_0 \ldots a_p$. Hence they may be solved so as to give the a's in terms of the calculable quantities $S(X)$, $S(X^2)$, $\ldots$ $S(X^{2p})$, $S(Y)$, $S(YX)$, $\ldots$ $S(YX^p)$.

17.12. It will be seen that the solution of these equations depends on the evaluation of the various summed quantities. A first step is therefore to calculate these sums, and this is done by a process very similar to that used in finding the moments of a distribution.

We can, in fact, express the equations in terms of moments. Dividing each equation by n, and remembering that $\mu_r' = \frac{1}{n}S(X^r)$, we have:

$$\left. \begin{array}{l} \dfrac{1}{n}S(Y) \quad -a_0 \quad -a_1\mu_1' \quad -a_2\mu_2' \quad - \ldots -a_p\mu_p' \quad =0 \\[2mm] \dfrac{1}{n}S(YX) -a_0\mu_1' -a_1\mu_2' -a_2\mu_3' \quad - \ldots -a_p\mu_{p+1}' =0 \\[1mm] \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\[1mm] \dfrac{1}{n}S(YX^p) -a_0\mu_p' -a_1\mu_{p+1}' - a_2\mu_{p+2}' - \ldots -a_p\mu_{2p}' =0 \end{array} \right\} \quad (17.9)$$

Equations for Fitting a Straight Line.

17.13. In the simplest case, that of a straight line, we have $p=1$, and the equations (17.9) become:

$$\left. \begin{array}{l} \dfrac{1}{n}S(Y) \ =a_0 + a_1\mu_1' \\[2mm] \dfrac{1}{n}S(YX) =a_0\mu_1' + a_1\mu_2' \end{array} \right\} \qquad . \qquad . \qquad . \quad (17.10)$$

In particular, if X and Y are measured about their means and hence are denoted by x, y, we have:

$$\mu_1 = 0$$
$$S(y) = 0$$

and hence, from (17.10),

$$a_0 = 0$$
$$a_1 = \frac{1}{n\mu_2}S(yx)$$

so that the fitted line is

$$y = x\frac{1}{n\mu_2}S(yx) \qquad . \qquad . \qquad . \qquad . \quad (17.11)$$

i.e. passes through the mean of X and Y. This is, in fact, the first regression equation of (11.6) (p. 209) in another form.

17.14. In equation (17.2) it is customary to call x the " independent " variable and y the " dependent " variable. In any given case it is, as a rule, possible to regard either of the variables under consideration as the independent variable, and the other as the dependent variable. We shall then get two expressions, one giving variable A in terms of variable B, the

other giving B in terms of A; and there will be two curves of closest fit, just as there are two regression lines in the theory of correlation.

These two curves are not, in general, the same, and the result sounds a little paradoxical until we examine how the two curves are derived. We have, in fact, two definitions of closest fit, one minimising residuals of the type $(A - a_0 - a_1 B - \ldots)^2$, the other minimising residuals of the type $(B - a_0' - a_1' A - \ldots)^2$. On *a priori* grounds there is nothing to choose between the two.

17.15. Which of the two forms we choose will depend in practice on a variety of circumstances. Sometimes one variable is clearly marked out as the independent variable. For example, in considering the way in which a population varies with time, it is almost inevitable to regard the former as dependent on the latter, and not *vice versa*. In other cases the choice is dictated by the purpose in view. For instance, in expressing the relationship between current and resistance in an electric circuit, an investigator would probably take as the independent variable that factor over which he had direct control. Frequently, however, there is no guide of this kind, and it may be necessary to ascertain both curves.[1]

Calculation.

17.16. The calculations necessary to fit a curve by the method of least squares fall into two stages. First of all, the sums of squares which appear in equation (17.8) must be found, or, what amounts to the same thing, the moments. To fit a curve of degree p it is necessary to find $2p$ sums of the type $S(X^k)$ and $p+1$ sums of the type $S(YX^k)$ (including $S(Y)$). The work is best carried out systematically after the manner of Chapter 9, and several devices considerably shorten the arithmetical labour.

(*a*) By a suitable choice of origin and unit we can often reduce the given values of X and Y to smaller numbers—a great help in calculating the higher powers and sums. For instance, if the values of Y were 625, 650, 675, 700, we could take an origin at $y = 625$, and a scale of one unit $= 25$, and our new values would then be 0, 1, 2, 3.

(*b*) If the values of the independent variable proceed by equal steps, and particularly if there is an odd number of them, the labour of calculation is enormously reduced. We shall consider this important case in some detail below (17.22).

When the various sums have been ascertained, the second stage, that of the solution of the equations (17.8), may be carried through. For a curve of degree p there are $p+1$ of these equations. They are linear in the unknowns a, and their solution offers only arithmetical difficulty.

17.17. Before proceeding to consider some examples, we may remark

---

[1] In this connection we may refer to a problem for which, so far as we are aware, no general solution has been found. Given that the theoretical law relating y and x is linear, but that the sets of values given in the data are *both* subject to error, what is the unique straight line most probably (in some sense) representing the truth? The least squares solutions will give us lines which, in a certain sense, are the most likely if the dependent variable is subject to errors normally distributed; but they do not yield a line which allows for errors in both variables.

Greenwood and Yule (*Proc. Roy. Soc. Medicine*, vol. 8, 1915, p. 113, Section of Epidemiology) used the principal axis (12.9) as an empirically good solution. This makes the sum of squares of perpendiculars from the points on to the line a minimum.

The difficulty is greatly intensified if the theoretical law is a polynomial of degree higher than the first.

on one point of theoretical interest. It is always possible to fit a curve of degree p exactly to $p+1$ points; for instance, a straight line can be drawn to pass exactly through two points, a cubic parabola through four points, and so on. Thus, if we have n points we can always find a curve of degree $n-1$ which is an exact fit. But in practice n is rarely less than ten, and a fitted curve of degree as high as this would have no practical value and very little theoretical interest. It is only exceptionally that use is found for fitted curves of degree higher than the fourth.

We will now consider some examples.

Example 17.1.—Let us fit a straight line to the data of Table 17.1. To illustrate the method we will deal with both cases, taking first distance and then velocity as the independent variable.

Denoting, then, distance by x and velocity by y, we wish to fit a curve of the form

$$y = a_0 + a_1 x$$

For this we require $S(X)$, $S(X^2)$, $S(Y)$ and $S(YX)$. For the alternative case we shall also require $S(Y^2)$.

The arithmetic is shown in Table 17.2. In successive columns we write, for each nebula, Y, X, X^2, YX and Y^2. Totals are shown at the foot of the columns.

Equations (17.8) then become:

$$S(Y) - a_0 n - a_1 S(X) = 0$$
$$S(YX) - a_0 S(X) - a_1 S(X^2) = 0$$

or

$$61 \cdot 26 - 10 a_0 - 114 \cdot 25 a_1 = 0$$
$$1261 \cdot 4988 - 114 \cdot 25 a_0 - 2371 \cdot 6145 a_1 = 0$$

Multiplying the first of these by 114·25 and the second by 10, and subtracting, we get

$$5616 \cdot 033 - 10,663 \cdot 0825 a_1 = 0$$
$$a_1 = 0 \cdot 527 \text{ (more accurately, } 0 \cdot 526,680,066)$$

and hence,

$$a_0 = 0 \cdot 109 \text{ (more accurately, } 0 \cdot 108,680,240).$$

So that

$$y = 0 \cdot 109 + 0 \cdot 527 x \qquad . \qquad . \qquad . \qquad . \qquad (a)$$

This line is shown in fig. 17.1.

If we wish to express distance in terms of velocity, we have, interchanging X and Y in equations (17.8):

$$x = a_0' + a_1' y$$
$$S(X) - a_0' n - a_1' S(Y) = 0$$
$$S(XY) - a_0' S(Y) - a_1' S(Y^2) = 0$$

or

$$114 \cdot 25 - 10 a_0' - 61 \cdot 26 a_1' = 0$$
$$1261 \cdot 4988 - 61 \cdot 26 a_0' - 672 \cdot 8998 a_1' = 0$$

whence

$$a_0' = -0 \cdot 135$$
$$a_1' = 1 \cdot 89$$

and

$$x = -0 \cdot 135 + 1 \cdot 89 y \qquad . \qquad . \qquad . \qquad . \qquad (b)$$

TABLE 17.2.—*Practical Work for Fitting a Straight Line to the Data of Table 17.1.*

Constellation.	Mean Velocity (000 km. per second). Y.	Distance (millions of parsecs). X.	X^2.	YX.	Y^2.
Isolated Nebula II .	0·63	1·20	1·4400	0·7560	0·3969
Virgo . .	0·89	1·82	3·3124	1·6198	0·7921
Isolated Nebula I .	2·35	3·31	10·9561	7·7785	5·5225
Pegasus . .	3·81	7·24	52·4176	27·5844	14·5161
Pisces . .	4·63	6·92	47·8864	32·0396	21·4369
Cancer . .	4·82	9·12	83·1744	43·9584	23·2324
Perseus . .	5·23	10·97	120·3409	57·3731	27·3529
Coma . .	7·50	14·45	208·8025	108·3750	56·2500
Ursa Major. .	11·80	22·91	524·8681	270·3380	139·2400
Leo . .	19·60	36·31	1318·4161	711·6760	384·1600
Total . .	61·26	114·25	2371·6145	1261·4988	672·8998

Equations (*a*) and (*b*) are nearly identical, for dividing (*a*) by 0·527 and rearranging, we have:

$$x = -0·207 + 1·90y$$

This is exceptional, and results from the closeness with which the points lie to a straight line. The correlation between X and Y is, in fact, 0·997.

Reduction of Data to Linear Form.

17.18. *Example 17.2.*—It sometimes happens that we may reduce data to a linear form by some simple transformation. Table 17.3, for example, shows the number of fronds of a duckweed plant on fourteen successive days. The number of fronds (N) clearly does not increase uniformly with time (x), and the curve of growth is not linear, as may be seen by graphing N against x. There are theoretical reasons for inquiring whether the law of growth may be represented by an equation of the form

$$N = ae^{bx}$$

A population which conformed to this equation would have the property that its rate of increase at any moment was proportional to the size of the population at that moment—its "birth-rate," so to speak, would be a constant.

Taking logarithms, we have:

$$\log_e N = \log_e a + bx$$

and if we now write $y = \log_e N$, we have:

$$y = \log_e a + bx$$

which is linear in x and y.

We should, of course, have a relation of the same form, with different values of the constants a and b, if we took logarithms to base 10, which is usually the more convenient procedure.

We therefore try the effect of fitting a straight line to x (the time) and

$\log_{10} N$ (log number of fronds). From fig. 17.2 it will be seen that the fit is a close one.
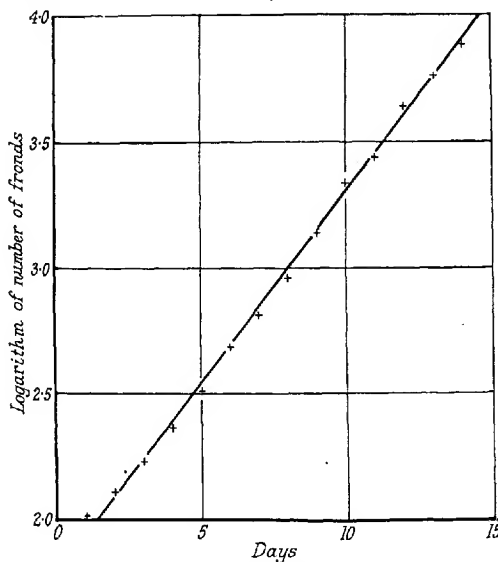
FIG. 17.2.—Straight Line fitted to Data of Table 17.3.
(Growth of Duckweed.)

TABLE 17.3.—*Growth of Duckweed.* (V. H. Blackman, *Nature*, 6th June 1936, quoting data of Ashby and Oxley.)

Number of Fronds. N.	$\log_{10} N$. Y.	Days. X.	X^2.	YX.
100	2·0000000	1	1	2·0000000
127	2·1038037	2	4	4·2076074
171	2·2329961	3	9	6·6989883
233	2·3673559	4	16	9·4694236
323	2·5092025	5	25	12·5460125
452	2·6551384	6	36	15·9308304
654	2·8155777	7	49	19·7090439
918	2·9628427	8	64	23·7027416
1406	3·1479853	9	81	28·3318677
2150	3·3324385	10	100	33·3243850
2800	3·4471580	11	121	37·9187380
4140	3·6170003	12	144	43·4040036
5760	3·7604225	13	169	48·8854925
8250	3·9164539	14	196	54·8303546
Total	40·8683755	105	1015	340·9594891

The preliminary work is shown in Table 17.3. We find first Y, corresponding to $\log_{10} N$, then $S(X)$, $S(Y)$, $S(X^2)$, $S(YX)$. For this particular example we do not require $S(Y^2)$. In view of the simple character of the values of X there is little saving in taking other origins or units for X and Y, although, if we were fitting a curve of higher order, it might be an advantage to take a different origin for X.

Equations (17.8) then become :

$$S(Y) - na_0 - a_1 S(X) = 0$$
$$S(YX) - a_0 S(X) - a_1 S(X^2) = 0$$

or

$$40 \cdot 8683755 - 14a_0 - 105a_1 = 0$$
$$340 \cdot 9594891 - 105a_0 - 1015a_1 = 0$$

whence

$$a_0 = 1 \cdot 785$$
$$a_1 = 0 \cdot 1514$$

and

$$y = 1 \cdot 785 + 0 \cdot 1514x \quad . \qquad . \qquad . \qquad . \qquad (a)$$

Raising this to power 10, and remembering that $10^y = N$, we have :

$$N = 10^{1 \cdot 785} \times 10^{0 \cdot 1514x} \quad . \qquad . \qquad . \qquad . \qquad (b)$$

which we may also write, expressing the powers of 10 as actual numbers :

$$N = 60 \cdot 95 \times (1 \cdot 417)^x$$

17.19. *Example 17.3.*—The process of taking logarithms may be applied to both variables. In Table 17.4 are given the costs per unit of electricity sold (η) and the number of units sold per head of the population served by the undertaking (ξ) for 27 electricity undertakings. The data were taken from the Returns of the Electricity Commission for 1933–34, which cover about six hundred undertakings, by selecting every twenty-fifth. They are, therefore, only a comparatively small sample, but they reflect fairly accurately the general relationship between ξ and η for the whole number of undertakings.

This relationship is illustrated by fig. 17.3, on which ξ is graphed against η. It will be seen that, broadly, the larger the number of units sold per head, the lower the cost per unit.

The points of fig. 17.3 lie, in fact, about a curve which suggests a relation of the form :

$$\eta = a\xi^{-b}$$

As ξ becomes larger, η becomes smaller, and as ξ tends to zero, η tends to infinity. Let us try to fit a curve of this kind to the data.

We have :

$$\log \eta = \log a - b \log \xi$$

and, putting

$$y = \log \eta, \qquad x = \log \xi$$

$$y = \log a - bx$$

which is linear. We therefore proceed to fit a straight line to $\log \eta$ and $\log \xi$.
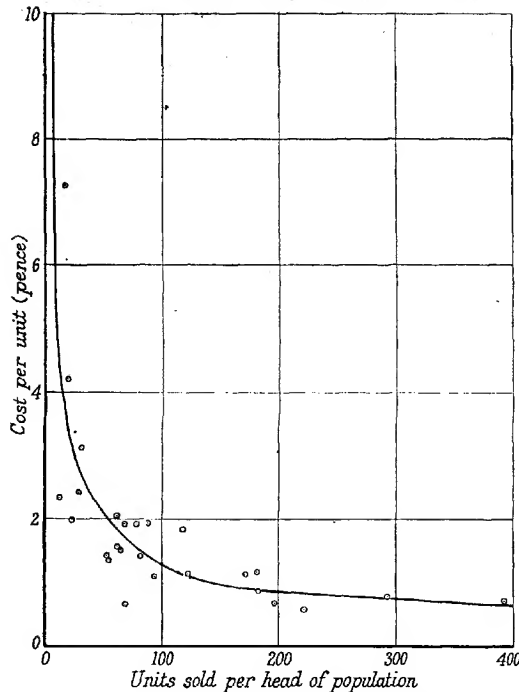
FIG. 17.3.—Curve fitted to Data of Table 17.4.

The preliminary work is shown in Table 17.4. Equations (17.8) become, in the usual way,

$$5\cdot2493 - 27a_0 - 50\cdot1311a_1 = 0$$

$$7\cdot3008 - 50\cdot1311a_0 - 97\cdot1450a_1 = 0$$

whence

$$a_0 = 1\cdot31 \qquad a_1 = -0\cdot601$$

and

$$y = 1\cdot31 - 0\cdot601x \qquad . \qquad . \qquad . \qquad . \qquad (a)$$

From which

$$\eta = 10^{1\cdot31}\xi^{-0\cdot601} \qquad . \qquad . \qquad . \qquad . \qquad (b)$$

or

$$\eta = 20\cdot42\xi^{-0\cdot601}$$

Fig. 17.4 shows the values of y plotted against those of x. The straight line we have found cannot be described as a good fit, but so far as the eye

TABLE 17.4.—REDUCTION OF NON-LINEAR RELATION TO LINEAR FORM: *Relationship between Working Costs per Unit and Number of Units Sold in 27 Electricity Undertakings.* (Data from Return of Engineering and Financial Statistics, 1933–34—Electricity Commission.)

Name of Undertaking.	Working Costs per Unit Sold (pence). η.	Units Sold (excluding bulk supplies) per Head of Population. ξ.	$\log \eta = Y.$	$\log \xi = X.$	$YX.$	$X^2.$
Aberdare	1·53	63·1	0·18469	1·8000	0·3324	3·2400
Barry U.D.C.	2·36	12·1	0·37291	1·0828	0·4038	1·1725
Bredbury and Romiley	0·70	394·2	−0·15490	2·5957	−0·4021	6·7377
Chesterfield	0·56	220·5	−0·25181	2·3434	−0·5901	5·4915
Earby	1·41	52·4	0·14922	1·7193	0·2566	2·9560
Grange	1·88	119·4	0·27416	2·0770	0·5694	4·3139
Holmfirth	1·17	181·6	0·06819	2·2591	0·1541	5·1035
Lincoln	0·78	293·8	−0·10791	2·4681	−0·2663	6·0915
Mexborough	1·13	170·4	0·05308	2·2315	0·1185	4·9796
Nuneaton	0·86	184·1	−0·06550	2·2651	−0·1484	5·1307
Redcar	1·91	68·0	0·28103	1·8325	0·5150	3·3581
Slaithwaite	1·40	80·7	0·14613	1·9069	0·2787	3·6363
Tanfield	2·41	29·0	0·38202	1·4624	0·5587	2·1386
West Lancs R.D.C.	1·37	53·4	0·13672	1·7275	0·2362	2·9843
Dumfries Corporation	1·10	93·0	0·04139	1·9685	0·0815	3·8750
Tobermory	4·21	19·9	0·62428	1·2989	0·8109	1·6871
Aberayron	8·9	25·6	0·94939	1·4082	1·3369	1·9830
Brixham Gas and Electric Co.	3·13	30·4	0·49554	1·4829	0·7348	2·1990
Chudleigh Co.	7·28	16·7	0·86213	1·2227	1·0541	1·4950
Foots Cray Co.	1·92	77·8	0·28330	1·8910	0·5357	3·5759
Lewes Co.	1·14	120·1	0·05690	2·0795	0·1183	4·3243
Newcastle Electric Light Co.	0·64	68·8	−0·19382	1·8376	−0·3562	3·3768
Ramsgate Co.	1·57	60·5	0·19590	1·7818	0·3490	3·1748
Steyning Co.	1·06	93·9	0·02531	1·9727	0·0499	3·8915
West Devon Co.	1·98	22·1	0·29667	1·3444	0·3988	1·8074
Coatbridge and Airdrie Co.	0·68	196·2	−0·16749	2·2927	−0·3840	5·2565
Skelmorlie Co.	2·05	60·1	0·31175	1·7789	0·5546	3·1645
Total	—	—	5·24928	50·1311	7·3008	97·1450

can judge it is as good as any simple curve is likely to be. It expresses the general relation between x and y; but, naturally, local circumstances cause individual values to deviate appreciably from this relation. Statistical data which are not produced under laboratory conditions are very often of this nature. The fitted curve expresses a general trend, but individual cases may lie well away from it in a number of instances.

Fitting of More General Curves.

17.20. *Example 17.4.*—We must now consider the fitting of curves of order higher than the first.

Table 17.5 shows the percentage loss of weight (Y) for certain temperatures (X) in experiments on the oven-drying of soils. Since X is

here the controllable factor, it is natural to take it as the independent variable, and we shall express Y in terms of X.
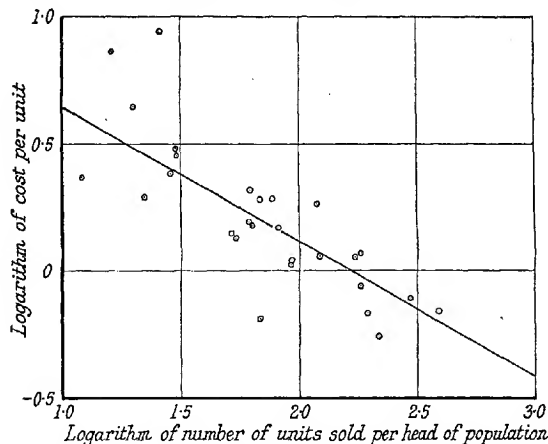
FIG. 17.4.—Straight Line fitted to Logarithms of Data of Table 17.4.

The data are shown graphically in fig. 17.5. We shall find successively the straight line, quadratic parabola and cubic parabola of closest fit. We shall therefore require sums of powers of X up to $S(X^6)$ and sums of products up to $S(YX^3)$. We also require, for later work, $S(Y^2)$.

The preliminary work is shown in Table 17.5. We might, perhaps, have abbreviated the arithmetic slightly by taking an origin of x at $X=100$ and of y at $Y=3$, but the saving would not have been large. Data of this kind frequently give rise to large figures in the higher sums, and a machine is a great help in the calculation. For instance, with a machine the sums $S(YX)$, etc., can be found by continuous addition, without the necessity for writing each individual contribution in the relative column.

For the straight line of closest fit, equations (17.8) become:

$$82 \cdot 97 - 16a_0 - 2642a_1 = 0$$

$$14{,}736 \cdot 19 - 2642a_0 - 474{,}050a_1 = 0$$

whence

$$a_0 = 0 \cdot 660 \quad \text{and} \quad a_1 = 0 \cdot 02741$$

(more accurately, $0 \cdot 659{,}759{,}789$ and $0 \cdot 027{,}408{,}722$)

and the straight line is:

$$y = 0 \cdot 660 + 0 \cdot 02741x \quad . \qquad . \qquad . \qquad . \qquad (a)$$

For the quadratic parabola, equations (17.8) are:

$$S(Y) \quad - na_0 \quad - a_1 S(X) \quad - a_2 S(X^2) = 0$$
$$S(YX) \quad - a_0 S(X) \quad - a_1 S(X^2) - a_2 S(X^3) = 0$$
$$S(YX^2) - a_0 S(X^2) - a_1 S(X^3) - a_2 S(X^4) = 0$$

21

TABLE 17.5.—*Curve-fitting to express the Relationship between Temperature and Percentage Loss in Weight of Certain Soil Samples.* (Data from J. R. H. Coutts, "'Single Value' Soil Properties: V. On the Changes Produced in a Soil by Oven-drying," *Journal Agricultural Science*, vol. 20, 1930, pp. 541–548.)

Percentage Loss in Weight. Y.	Temperature (degrees). X.	Y².	X².	X³.	X⁴.	X⁵.	X⁶.	YX.	YX².	YX³.
3·71	100	13·7641	10,000	1,000,000	100,000,000	10,000,000,000	1,000,000,000,000	371·00	37,100·00	3,710,000·00
3·81	105	14·5161	11,025	1,157,625	121,550,625	12,762,815,625	1,340,095,640,625	400·05	42,005·25	4,410,551·25
3·86	110	14·8996	12,100	1,331,000	146,410,000	16,105,100,000	1,771,561,000,000	424·60	46,706·00	5,137,760·00
3·93	115	15·4449	13,225	1,520,875	174,900,625	20,113,571,875	2,313,060,765,625	451·95	51,974·25	5,977,038·75
3·96	121	15·6816	14,641	1,771,561	214,358,881	25,937,424,601	3,138,428,376,721	479·16	57,978·36	7,015,381·56
4·20	132	17·6400	17,424	2,299,968	303,595,776	40,074,642,432	5,289,852,801,024	554·40	73,180·80	9,659,865·60
4·34	144	18·8356	20,736	2,985,984	429,981,696	61,917,364,224	8,916,100,448,256	624·96	89,994·24	12,959,170·56
4·51	153	20·3401	23,409	3,581,577	547,981,281	83,841,135,993	12,827,693,806,929	690·03	105,574·59	16,152,912·27
4·73	163	22·3729	26,569	4,330,747	705,911,761	115,063,617,043	18,755,369,578,009	770·99	125,671·37	20,484,433·31
5·35	179	28·6225	32,041	5,735,339	1,026,625,681	183,765,996,899	32,894,113,444,921	957·65	171,419·35	30,684,063·65
5·74	191	32·9476	36,481	6,967,871	1,330,863,361	254,194,901,951	48,551,226,272,641	1,096·34	209,400·94	39,995,579·54
6·14	203	37·6996	41,209	8,365,427	1,698,181,681	344,730,881,243	69,980,368,892,329	1,246·42	253,023·26	51,363,721·78
6·51	212	42·3801	44,944	9,528,128	2,019,963,136	428,232,184,832	90,785,223,184,384	1,380·12	292,585·44	62,028,113·28
6·98	226	48·7204	51,076	11,543,176	2,608,757,776	589,579,257,376	133,244,912,279,976	1,577·48	356,510·48	80,671,368·48
7·44	237	55·3536	56,169	13,312,053	3,154,956,561	747,724,704,957	177,210,755,074,809	1,763·28	417,897·36	99,041,674·32
7·76	251	60·2176	63,001	15,813,251	3,969,126,001	996,250,626,231	250,058,907,189,001	1,947·76	488,887·76	122,710,827·76
82·97	2642	439·4363	474,060	91,244,582	18,553,164,842	3,930,294,225,302	858,077,668,755,250	14,736·19	2,819,909·45	571,902,362·11

These become, on substitution,

$$82 \cdot 97 - 16a_0 - 2642a_1 - 474{,}050a_2 = 0$$

$$14{,}736 \cdot 19 - 2642a_0 - 474{,}050a_1 - 91{,}244{,}582a_2 = 0$$

$$2{,}819{,}909 \cdot 45 - 474{,}050a_0 - 91{,}244{,}582a_1 - 18{,}553{,}164{,}842a_2 = 0$$

giving

$$a_0 = 3 \cdot 551, \qquad a_1 = -0 \cdot 009291, \qquad a_2 = 0 \cdot 00010695$$

(more accurately, $3 \cdot 550{,}990{,}2$, $-0 \cdot 009{,}291{,}235{,}7$, and $0 \cdot 000{,}106{,}954{,}12$)

and the parabola is:

$$y = 3 \cdot 551 - 0 \cdot 009291x + 0 \cdot 00010695x^2 \qquad . \qquad . \qquad (b)$$

For the cubic parabola, equations (17.8) are :—

$$S(Y) \quad - na_0 \quad - a_1 S(X) \quad - a_2 S(X^2) - a_3 S(X^3) = 0$$
$$S(YX) \quad - a_0 S(X) \quad - a_1 S(X^2) - a_2 S(X^3) - a_3 S(X^4) = 0$$
$$S(YX^2) - a_0 S(X^2) - a_1 S(X^3) - a_2 S(X^4) - a_3 S(X^5) = 0$$
$$S(YX^3) - a_0 S(X^3) - a_1 S(X^4) - a_2 S(X^5) - a_3 S(X^6) = 0$$

which become :

$$82 \cdot 97 - 16a_0 - 2642a_1 - 474{,}050a_2 - 91{,}244{,}582a_3 = 0$$
$$14{,}736 \cdot 19 - 2642a_0 - 474{,}050a_1 - 91{,}244{,}582a_2 - 18{,}553{,}164{,}842a_3 = 0$$
$$2{,}819{,}909 \cdot 45 - 474{,}050a_0 - 91{,}244{,}582a_1 - 18{,}553{,}164{,}842a_2 - 3{,}930{,}294{,}225{,}302a_3 = 0$$
$$571{,}902{,}362 \cdot 11 - 91{,}244{,}582a_0 - 18{,}553{,}164{,}842a_1 - 3{,}930{,}294{,}225{,}302a_2 - 858{,}077{,}668{,}755{,}250a_3 =$$

It is not really necessary to write out the large numbers of the later equations as fully as we have done, and a certain amount of approximation is allowable. The student should, however, be careful not to introduce it too soon, as neglected quantities may become of cumulative importance in the solution of the equations.

By straightforward but rather strenuous arithmetic we find :

$$a_0 = 7 \cdot 783, \qquad a_1 = -0 \cdot 08940$$
$$a_2 = 0 \cdot 0005875, \qquad a_3 = -0 \cdot 0000009189$$

(more accurately, $a_0 = 7 \cdot 782{,}526{,}861$, $a_1 = -0 \cdot 089{,}402{,}395{,}60$

$a_2 = 0 \cdot 000{,}587{,}479{,}234{,}2$, $a_3 = -0 \cdot 000{,}000{,}918{,}891{,}069{,}8$)

The smallness of the coefficients a_2 and a_3 does not mean that they are of minor importance, since in the equation for y they are multiplied by terms in x^2 and x^3, which may be large.

The cubic parabola is, then,

$$y = 7 \cdot 783 - 0 \cdot 08940x + 0 \cdot 0005875x^2 - 0 \cdot 0000009189x^3$$

which we may also write as :

$$y = 7 \cdot 783 - 8 \cdot 940 \frac{x}{100} - 5 \cdot 875 \left(\frac{x}{100}\right)^2 - 0 \cdot 9189 \left(\frac{x}{100}\right)^3 \qquad . \qquad (c)$$

Fig. 17.5 shows the data graphically, with the straight line and cubic parabola of closest fit.
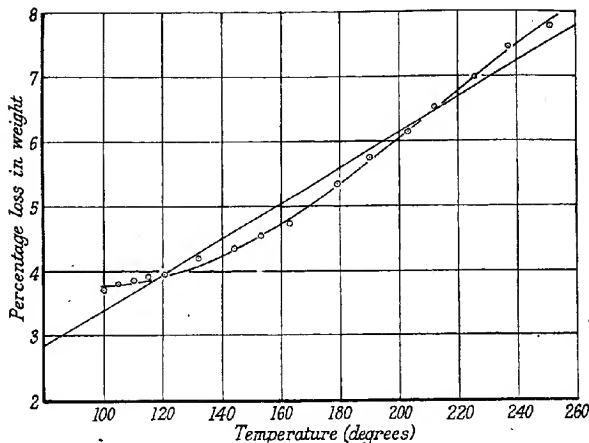
FIG. 17.5.—Straight Line and Cubic Parabola of Closest Fit to the
Data of Table 17.5.

17.21. Although a graph will usually suggest whether a straight line
or quadratic parabola is likely to give a satisfactory fit, it will not as a rule
be much guide in deciding whether further terms will repay the labour
of calculation. This can be judged, at least roughly, by calculating
the terms given by the polynomial (to as high a degree as it has been
carried) for the observed values of x, and then observing the run of the
residuals. If the signs run more or less at random it will hardly be
worth while to calculate another term ; but if a series of positive residuals
is followed by a series of negative residuals, these by another series of
positive residuals, etc.. it will probably be worth while to proceed further.
Moreover, the coefficients for a parabola of order k are no guide to those
of order $k+1$. For instance, in Example 17.4, the values of a_0 for the
straight line, square parabola and cubic parabola are 0·660, 3·551, 7·783 ;
and those of a_1 are 0·02741, − 0·009291, − 0·08940. From this informa-
tion we could not guess even the sign of these coefficients in the parabola
of order 4, and if we wished to fit such a curve five equations of the type
(17.8) would have to be solved *ab initio*.

The student, therefore, should not fall into the error of thinking that
parabolas of successive orders will resemble each other in their lower
terms, or that the fitting of a curve of order $k+1$ is merely a question of
adding an extra term to a curve of order k. It would be a great con-
venience if this were so, and, in fact, methods have been devised whereby
one variate can be expressed in terms of certain polynomials of the other
in such a way that this advantage is secured. The theory of these
so-called " orthogonal " polynomials is, however, outside the scope of
the present work, and we would refer the student who is interested to
the references for this chapter.

The Case when the Independent Variable Proceeds by Equal Steps.

17.22. When the independent variable x proceeds by steps of equal amount h, the arithmetical solution of equations (17.8) can be greatly simplified, particularly if the number of values is odd. In such a case we take h as the unit of x and an origin at the middle term. The values of x will then be $-k$, $-(k-1)$, $-(k-2)$, . . . -2, -1, 0, 1, 2, . . . $(k-2)$, $(k-1)$, k, and owing to the symmetry of this series the sums of odd powers of x will vanish, *i.e.* $S(X)$, $S(X^3)$, $S(X^5)$, etc. are all zero. Equations (17.8) then become, taking p as odd,

$$
\begin{aligned}
S(Y) \quad &- na_0 &&- a_2 S(X^2) &&- a_4 S(X^4) \ . \ . \ . &&= 0\\
S(YX) \quad &&- a_1 S(X^2) &&- a_3 S(X^4) \ . \ . \ . &&= 0\\
& \quad . &&\quad . &&\quad . &&\quad . \\
S(YX^{p-1}) &- a_0 S(X^{p-1}) &&- a_2 S(X^{p+1}) \ . \ . \ . &&&&= 0\\
S(YX^p) \quad &&- a_1 S(X^{p+1}) &&- a_3 S(X^{p+3}) \ . \ . \ . &&= 0
\end{aligned}
\qquad (17.12)
$$

and not only is the number of terms reduced, but the equations split into two sets, one in a_0, a_2, a_4, etc., and the other in a_1, a_3, a_5, etc. Moreover, the sums of even powers of X are twice the sums of powers of the first k natural numbers, which may be easily found, either from tables or from known formulæ.

Example 17.5.—Table 17.6 shows the population of England and Wales in certain census years from 1811 onwards. Taking the time as the independent variable, we choose as the unit of X the period of ten years, and the origin at the mid-point of the range, 1871. The preliminary work for the fitting of curves up to the cubic form is shown in the table.

For the cubic parabola, equations (17.8) are, then,

$$
\begin{aligned}
314 \cdot 09 - 13a_0 \quad &&- 182a_2 && &&= 0\\
474 \cdot 77 \quad &- 182a_1 && &&- 4550a_3 &&= 0\\
4520 \cdot 45 - 182a_0 \quad && - 4550a_2 && &&= 0\\
11{,}632 \cdot 97 \quad &- 4550a_1 && &&- 134{,}342a_3 &&= 0
\end{aligned}
$$

whence

$$
\begin{aligned}
a_0 &= 23 \cdot 299 & a_1 &= \ 2 \cdot 895\\
a_2 &= \ 0 \cdot 06153 & a_3 &= -0 \cdot 01147
\end{aligned}
$$

The parabola is, therefore,

$$
y = 23 \cdot 299 + 2 \cdot 895x + 0 \cdot 06153x^2 - 0 \cdot 01147x^3 \quad . \qquad . \qquad (a)
$$

Fig. 17.6 shows the data graphically, together with this cubic.

Incidentally, this example illustrates one point of some importance. Over the years 1811 to 1931 the cubic gives a fair fit, and might be used to estimate the population at intermediate years. But for extrapolation it is of very little value. We could not estimate the population for 1951 with any confidence by putting $x = 8$ in the cubic; still less that for later years. Unless there are good reasons for supposing that the fitted curve is an accurate representation of a theoretical relationship, it is dangerous

to assume that a fitted parabola can be used outside the range for which it was ascertained.

TABLE 17.6.—*Curve-fitting to Growth of Population in England and Wales.* (Data from Registrar-General's Statistical Review of England and Wales, 1933, Tables, Part II.)

Year.	Population (millions) Y.	X.	X².	X³.	X⁴.	X⁶.	YX.	YX².	YX³.
1811	10·16	− 6	36	− 216	1,296	46,656	− 60·96	365·76	− 2,194·56
1821	12·00	− 5	25	− 125	625	15,625	− 60·00	300·00	− 1,500·00
1831	13·90	−·4	16	− 64	256	4,096	− 55·60	222·40	− 889·60
1841	15·91	− 3	9	− 27	81	729	− 47·73	143·19	− 429·57
1851	17·93	− 2	4	− 8	16	64	− 35·86	71·72	− 143·44
1861	20·07	− 1	1	− 1	1	1	− 20·07	20·07	− 20·07
1871	22·71	0	0	0	—	—	—	—	—
1881	25·97	1	1	1	1	1	25·97	25·97	25·97
1891	29·00	2	4	8	16	64	58·00	116·00	232·00
1901	32·53	3	9	27	81	729	97·59	292·77	878·31
1911	36·07	4	16	64	256	4,096	144·28	577·12	2,308·48
1921	37·89	5	25	125	625	15,625	189·45	947·25	4,736·25
1931	39·95	6	36	216	1,296	46,656	239·70	1,438·20	8,629·20
Total	314·09	0	182	0	4,550	134,342	474·77	4,520·45	11,632·97
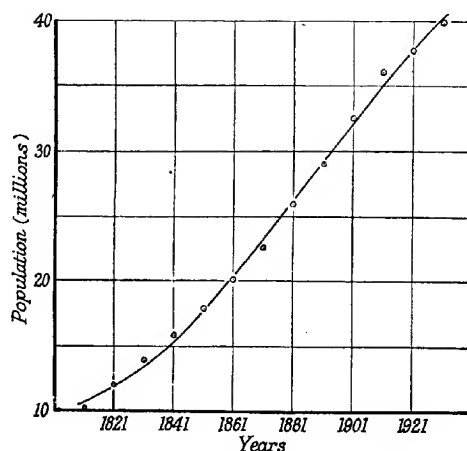
FIG. 17.6.—Cubic Parabola fitted to the Data of Table 17.6.

It would be instructive for the student to fit merely a segment of some actual series and note how rapidly the curve calculated from the segment diverged from the observations outside its limits. It has been shown that even within the limits of the fitted observations the fit tends to be worst

as the limits are approached. The higher powers of x become of greater, and greater effect the more we diverge from the centre of the fitted segment and tend, so to speak, to " wag the tail " of the curve.

17.23. If the number of values of x is even, we have a choice of two methods of procedure. We can take h as unit and the origin at one of the two middle values; or we can take $\frac{1}{2}h$ as unit and origin midway between the two central values. In the first case, the sums of odd powers will no longer vanish, but they will nevertheless be easily calculable, since all terms except a single outlying member in the summation will cancel out in pairs. In the second case the sums of odd powers will vanish, but the other sums will no longer be twice those of the first k natural numbers, but of the first k odd numbers. In either case the solution of the equations (17.8) is not difficult.

Calculation of the Sum of Squares of Residuals.

17.24. The eye is not a reliable guide to the closeness with which a given curve lies to data, and it is desirable to have some more accurate measure of the closeness of fit. For this purpose we require to be able to find the sum of the squares of residuals U. We know by our method of ascertaining the curve that this will be less than the corresponding quantity for any other curve of the same degree, and our interest is centred on how close this is to the ideal value zero.

To calculate the sum of squares of residuals it is not necessary to calculate each separate residual. In fact, for the parabola of order p we have :

$$U = S(Y - a_0 - a_1 X - a_2 X^2 - \ldots - a_p X^p)^2$$
$$= S\{Y(Y - a_0 - a_1 X - \ldots - a_p X^p)\}$$

for the terms of the type $S\{a_k X^k (Y - a_0 - a_1 X - \ldots - a_p X^p)\}$ vanish in virtue of equations (17.8). Hence,

$$U = S(Y^2) - a_0 S(Y) - a_1 S(YX) - \ldots - a_p S(YX^p) \qquad (17.13)$$

The constants a and the sums which appear in this expression have already been found, with the exception of $S(Y^2)$ in some cases. With this additional quantity we can find U.

Example 17.6.—Let us find U for the data of Example 17.4 for the straight line and the two parabolas.

For the line
$$U = S(Y^2) - a_0 S(Y) - a_1 S(YX)$$

Here
$$S(Y) = 82 \cdot 97, \quad S(YX) = 14{,}736 \cdot 19$$
$$S(Y^2) = 459 \cdot 4363, \qquad a_0 = 0 \cdot 659{,}759{,}789$$
$$a_1 = 0 \cdot 027{,}408{,}722$$

Hence,
$$U = 459 \cdot 4363 - 54 \cdot 74027 - 403 \cdot 90014$$
$$= 0 \cdot 7959$$

For the quadratic parabola :

$$U = S(Y^2) - a_0 S(Y) - a_1 S(YX) - a_2 S(YX^2)$$

and here

$$a_0 = 3 \cdot 550,990,2$$
$$a_1 = -0 \cdot 009,291,235,7$$
$$a_2 = 0 \cdot 000,106,954,12$$

whence

$$U = 0 \cdot 1271$$

Similarly, for the cubic

$$U = 0 \cdot 0485$$

The value of U therefore decreases from $0 \cdot 7959$ for the straight line to $0 \cdot 0485$ for the cubic. This is what we should expect, for the addition of extra terms means that we have additional constants at our disposal in the task of minimising U.

To obtain U with any accuracy by the foregoing method it is necessary to ascertain the a's to a considerable number of decimal places.

Measurement of the Closeness of Fit.

17.25. The value of U enables us to make some sort of comparison between the fits of different curves to the same data ; but it is not, in itself, a satisfactory measure of fit, since it does not permit of the comparison of the fits of curves to different data. The measure U/n, which is the variance of errors of estimation, suggests itself, but this, like U, is not absolute, being dependent on the units in which we are working. For a satisfactory measure some form of ratio would have to be taken.

Such a ratio arises in a natural way if we consider the correlation between the actual values of Y and those " predicted " by the polynomial.

Let us, without loss of generality, suppose that the values are measured from their mean, and let y_r be the value given by the polynomial and Y_r be the actual value. Then, as in 17.24,

$$S(y^2) = S(Yy) \qquad . \qquad . \qquad . \qquad . \quad (17.14)$$

$$U = S\{Y(Y - y)\}$$
$$= S(Y^2) - S(Yy) \qquad . \qquad . \qquad . \quad (17.15)$$

Writing σ_Y, σ_y for the standard deviations of Y and y, and R for the correlation between them, we get, from (17.14),

$$\sigma_y^2 = R \sigma_Y \sigma_y$$

or

$$\sigma_y = R \sigma_Y \qquad . \qquad . \qquad . \qquad . \quad (17.16)$$

and from (17.15),

$$\frac{U}{n} = \sigma_Y^2 - R \sigma_Y \sigma_y$$

or

$$R \frac{\sigma_y}{\sigma_Y} = 1 - \frac{U}{n \sigma_Y^2} \qquad . \qquad . \qquad . \qquad . \quad (17.17)$$

Hence, substituting for σ_y from (17.16),

$$R^2 = 1 - \frac{U}{n\sigma_Y{}^2} \qquad . \qquad . \qquad . \qquad (17.18)$$

which gives the correlation in terms of the ratio of U/n and the variance $\sigma_Y{}^2$.

R is, in fact, analogous to the multiple correlation coefficient and the correlation ratio, and the equation (17.18) should be compared with equation (13.3), page 244, and equation (14.15), page 278.

Example 17.7.—In Example 17.1 we have, using the data of Table 17.2 and the constants found :

$$\sigma_Y{}^2 = 67 \cdot 28998 - (6 \cdot 126)^2$$
$$= 29 \cdot 762,104$$
$$U = 1 \cdot 835,777,255$$
$$R^2 = 1 - \frac{1 \cdot 835,777,255}{297 \cdot 62104} = 0 \cdot 993,831,830$$
$$R = 0 \cdot 99691$$

For the soil data of Examples 17.4 and 17.6 we find :

For the straight line $R = 0 \cdot 98627$
For the cubic $\qquad R = 0 \cdot 99917$

Thus, judged by the value of R, the straight line of Example 17.1 is a better fit than that of Example 17.4, but a worse fit than the cubic of the latter.

17.26. As a general comment on the scope of the methods of curve-fitting described in this chapter, we may remark that although polynomials can always be fitted to data, the student should not assume that even the polynomial of closest fit will necessarily be a *satisfactory* fit. It may exhibit peculiarities of behaviour which are entirely absent from the data themselves. He may well ask, when confronted by a given set of data, how he is to know whether they may be satisfactorily represented by a polynomial. The answer is that he must fit one and see. Some further remarks on this point are given later in **24.12**, where similar questions arise in connection with interpolation and graduation.

SUMMARY.

1. A parabola of the form $y = a_0 + a_1 x + a_2 x^2 + \ldots + a_p x^p$ may be fitted to data by choosing the constants a so that the sum of squares of residuals $U = S(Y - a_0 - a_1 X - a_2 X^2 - \ldots - a_p X^p)^2$ is a minimum.

2. This method leads to the equations

$$S(Y) \quad -na_0 \quad -a_1 S(X) \quad -a_2 S(X^2) \quad - \ldots -a_p S(X^p) \quad =0$$
$$S(YX) \quad -a_0 S(X) \quad -a_1 S(X^2) \quad -a_2 S(X^3) \quad - \ldots -a_p S(X^{p+1}) =0$$
$$S(YX^p) - a_0 S(X^p) - a_1 S(X^{p+1}) - a_2 S(X^{p+2}) - \ldots -a_p S(X^{2p}) \quad =0$$

3. Non-linear data may sometimes be reduced to the linear form by a simple transformation of one or both the variables.

4. The sum of squares of residuals may be found from the formula

$$U = S(Y^2) - a_0 S(Y) - a_1 S(YX) - \ldots - a_p S(YX^p)$$

5. One measure of the goodness of fit of the parabola to the data is given by R, the correlation between actual and " predicted " values of the variate. R is given by

$$R^2 = 1 - \frac{U}{n\sigma_Y^2}$$

where Y is the dependent variable.

EXERCISES.

17.1. Fit a straight line and parabolas of the second and third orders to the following data, taking X to be the independent variable—

X.	Y.
0	1
1	1·8
2	1·3
3	2·5
4	6·3

and find the sum of squares of residuals in the three cases.

17.2. (Data quoted by P. L. Fegiz, "Le variazioni stagionali della natalità," *Metron*, vol. 5, 1925, No. 4, p. 127.) The following figures show the relation between duration of marriage and average number of children per marriage in Norway in 1920 :—

Duration of Marriage (Years).	Average Number of Children.
0– 1	0·48
5– 6	2·09
10–11	3·26
15–16	4·33
20–21	5·14
25–26	5·63
30–31	5·77

By the method of least squares find equations of the first, second and third orders expressing the number of children in terms of the duration of marriage. Compare the values given by these expressions for a duration of 17–18 years with the true value 4·67.

17.3. The pressure of a gas and its volume are known to be related by an equation of the form $pv^\gamma = \text{constant}$.

In a certain experiment the following volumes of a quantity of the gas were observed for the pressures specified. Find the value of γ by fitting a straight line to the logarithms of p and v, taking p to be the independent variable.

p (kg. per square cm.) .	0·5	1·0	1·5	2·0	2·5	3·0
v (litres) . . .	1·62	1·00	0·75	0·62	0·52	0·46

17·4. (Data from the records of the Farm Economics Branch, School of Agriculture, Cambridge, England.)

The following are the gross output and the gross output per £100 of labour employed, for a selected number of farms :—

Gross Output (Units).	Gross Output per £100 Labour (Units).
63	40
223	155
755	188
165	78
1535	315
3193	290
2238	259
1228	231
2695	255

Fit a quadratic parabola to these data, taking gross output as the independent variable.

CHAPTER 18.

PRELIMINARY NOTIONS ON SAMPLING.

The Problem.

18.1. In practical problems the statistician is often confronted with the necessity of discussing a universe of which he cannot examine every member. For example, an inquirer into the heights of the population of Great Britain cannot afford the time or expense required to measure the height of each individual ; nor can a farmer who wants to know what proportion of his potato crop is diseased examine every single potato.

In such cases the best an investigator can do is to examine a limited number of individuals and hope that they will tell him, with reasonable trustworthiness, as much as he wants to know about the universe from which they come. We are thus led naturally to the question, What can be said about a universe of which we can examine only a limited number of its members ? This question is the origin of the **Theory of Sampling.**

18.2. A sample from a universe is a selected number of individuals each of which is a member of the universe. As a very special case the sample may consist of the entire universe.

It is a matter of common belief, founded on experience and intuition, that a sample will tell us something about the parent universe. The corn merchant, whose livelihood depends on his ability to ascertain the quality of the grain which he handles, is content to assess it by thrusting a conical trowel into the middle of a sack and scrutinising the sample he gets. He believes that the sample will be representative of the whole, and experience justifies him. He buys and sells on the basis of judgment from samples.

It is also a matter of common belief that the larger a sample becomes the more likely it is to reflect accurately the conditions in the parent universe.

To these and similar beliefs the theory of sampling gives a logical basis and a system of quantitative measurement. In this chapter we give a general survey of the fundamental ideas and the technique of sampling. In later chapters we shall develop these ideas and discuss their applications in various fields.

Types of Universe.

18.3. Before we consider sampling itself, however, it is desirable to look a little closer into the various types of universe which we shall have to investigate.

By a **finite universe** we shall mean a universe which contains a finite number of members. Such, for instance, is the universe of inhabitants of Great Britain and the universe of books in the British Museum.

Similarly, by an **infinite universe** we shall mean a universe containing an infinite number of members. Such, for instance, is the universe of pressures at various points in the atmosphere, or the universe of possible sizes of the wheat crop in tons, for, although there are limits to the size, the actual tonnage can take any numerical value within those limits.

. In many cases the number of members in a universe is so large as to be practically infinite. Moreover, a theoretical discussion of an infinite universe is frequently easier than a discussion of a finite universe, and a large class of problems may be treated by assuming that the parent universe is infinite, without introducing any sensible error.

It may be worth remarking that in a few cases we may be ignorant whether or not the universe of discussion is infinite. The universe of stars is an example.

Existent and Hypothetical Universes.

18.4. By the logical extension of the idea of a universe of concrete objects, which we shall call an **existent universe,** we are able to construct the idea of a **hypothetical universe.**

Consider the throws of a die. Each throw will be regarded as an individual. There is an infinite number of throws which can be made with the die, provided that it does not wear out. Let us then define as our universe of discussion all the *possible* throws of the die.

In doing so we are clearly making some new step ; for our universe is to be conceived as having no existence in reality but only in imagination. We can give actuality to some members of the universe by throwing the die, but we can never produce them all. Even if the die were locked away in a safe and never thrown at all there would still be a universe of possible throws.

Such a universe is called a **hypothetical** universe. We may define it formally as the aggregate of all the conceivable ways in which a specified event can happen. Other examples of hypothetical universes are the universe of all values which the bank rate can have in ten years' time, and the universe of the possible ways in which three balls can be arranged on a billiard table.

18.5. A hypothetical universe may, in fact, be imagined around any observed event. We have only to picture all the circumstances before the event happens ; the universe is then all the possible ways in which it could happen. Which of the ways it *will* happen does not affect the universe. We know that " from the chaos of predestination and the night of our forebeing " some one individual will emerge to assume the mantle of reality ; but which one that will be is another and more difficult question.

18.6. The student of metaphysics would perhaps criticise the thoughts expressed briefly in the previous two sections, but we have no space to go further into the philosophical implications of the idea of hypothetical universes. The problems which arise in this connection have, however, far more than an abstract interest. They lie at the root of a great many practical statistical problems, and most students, however utilitarian their outlook, will find that a clear perception of the issues involved may save a lot of thought and labour at a subsequent stage.

The literature on this subject, unfortunately, is scattered; but reference may with advantage be made to the works cited in refs. (888)–(890).

Universe of Universes.

18.7. Just as a universe may contain a number of sub-universes, so any given universe may be a member of some more widely defined universe. For example, the universe of inhabitants of Great Britain is a member of the universe of universes, each of which consists of the inhabitants of some European country.

Similarly, any existent universe may be regarded as one member of a hypothetical universe of universes. For instance, the normal universe of men whose heights have a mean of 65 inches and standard deviation 3 inches is a member of the hypothetical universe of all populations which are normally distributed with respect to height.

18.8. We shall sometimes have to discuss aggregates which it is difficult to regard as composed of individual members at all—for example, we may wish to sample a reservoir of water to test for pollution. In theory, perhaps, we could in such a case regard the reservoir as a universe composed of molecules each of which was an individual, but in practice, as we shall see, this is not usually a convenient method of approach. Such universes may frequently be treated as composed of arbitrary units, e.g. the reservoir may be regarded as composed of so many pints of fluid. Similarly, a 280-lb. sack of flour may be regarded as composed of 4480 ounces, and we can, if we like, regard it as weighed out into one-ounce packets.

18.9. We can now turn to discuss the aims which usually underlie a sampling inquiry.

Briefly, the fundamental object of sampling is to give the maximum information about the parent universe with the minimum effort. We must, therefore, consider the type of information we require and the methods by which it is to be obtained.

18.10. In sampling a universe we usually have in mind one or more of its variates. For instance, when we sample the population of Great Britain, we are not so much interested in the individuals as human beings as in one of their qualities, such as height or weight, or perhaps the correlation between height and weight. Our object will then be to get, from the sample, an idea of the frequency-distribution in the parent universe according to the chosen variates.

The ideal for this purpose would be to express the distribution in some mathematical form such as a Pearson curve (10.48). It may be, however, that the parent universe will not admit of this representation, or that the sample is not large enough for us to venture on it with any confidence.

In such cases we attempt to find estimates of certain constants of the parent universe. Very often this is all we need. We can, for example, form a very fair idea of the height distribution of the population of Great Britain if we know the mean and the standard deviation. If we can go further, and find the third and fourth moments, our idea will be better still.

Theory of Estimation.

18.11. Hence, a large part of the theory of sampling is devoted to finding from the sample estimates of certain constants of the parent

universe. Such constants include the measures of position and of dispersion, together with the moments and measures of skewness; and, in multivariate universes, the various total and partial correlations.

In general, there are more ways than one of estimating a constant from the data of the sample. Some of these ways will be better than others. The **Theory of Estimation** treats of these and cognate matters. It seeks to investigate the conditions which an estimate should obey, what are the best estimates to employ in given circumstances, and how good other estimates are in comparison.

Precision of Estimates.

18.12. It will be obvious that knowledge derived from a sample is not of the categorical kind customary in mathematics. If we have 1000 balls in a bag and draw 999 of them which turn out to be black, it is always *possible* that the remaining one is of some other colour. It is, however, so improbable, that in most practical cases we should be justified in concluding that the balls were all black.

If we did draw such a conclusion, and acted upon it, we should be basing our action, not upon certainty, but on probability. One does this kind of thing, of course, in nearly all everyday actions almost without noticing it. Some events, such as the death of a man before reaching the age of 150, have such a high degree of probability that we never regard them as other than certain ; other events, such as the possibility of rain to-morrow, are so uncertain that we should hesitate to make an important decision contingent upon them.

18.13. The second aim of the theory of sampling is, therefore, to determine as objectively as possible what degree of confidence we can put in our estimates when they are obtained. This we do in terms of probability as far as we can ; if this proves impossible, we sometimes have to rely on intuitive impressions or the results of previous experience, which are not expressible in quantitative terms.

Put in another way, we may say that our object is to determine the **precision** of an estimate. We attempt to do this by assigning limits to the probable divergence between the estimate based on the sample and the true value of the estimated quantity in the universe.

18.14. The accuracy of the estimate will depend on (*a*) the way in which the estimate is made from the data of the sample, and (*b*) the way in which the sample was obtained. Consideration of the first leads us again to the theory of estimation. The second leads us to study the **technique of sampling** and the **design of statistical inquiries.**

Tests of Significance.

18.15. If the sample is small we cannot, as a rule, assign to the estimates we obtain sufficiently narrow limits to locate the universe value with any serviceable accuracy. For example, a correlation of $+0.5$ in a sample of twelve might arise, rather infrequently, from a normal universe in which the true correlation was as high as $+0.9$ or as low as zero. For such samples our questions are accordingly framed in more qualitative terms : we do not ask, " What is the value of the correlation in the universe ? " but, " Is the observed value *significant* of the existence of any correlation at all in the universe, whatever its value ? " In other words,

we wish to know whether the observed value could have arisen from a universe in which the true correlation is zero. If our conclusion is that it could not, we may say that the sample value is significant of correlation, although 'we cannot say with much confidence what that correlation is.

Much of the investigation arising out of small samples is thus of a rather special character, and deals with **tests of significance**. The methods developed for the purpose of conducting such tests can be, and not infrequently are, applied also to large samples, either alone or supplementary to the direct approach of forming more or less precise estimates of the various quantities which specify the parent universe.

Types of Sampling.

18.16. The process of forming a sample consists of choosing a predetermined number of individuals from the parent universe. The choice may be exercised in three ways :

(a) By selecting the individuals at random (the meaning of " random " is discussed below).

(b) By selecting the individuals according to some purposive principle.

(c) By a mixture of (a) and (b).

Thus, in taking a sample of the inhabitants of Great Britain to study their income we might, according to method (a), select the individuals at random from census returns; or according to (b) we might, knowing roughly the average incomes in various age-groups, purposely select from each group an individual whose income was somewhere near the average in that group ; or (c) we might decide to take ten individuals from each group and select those ten by method (a).

18.17. Sampling of type (a) is called **random sampling**. That of type (b) is called **purposive sampling**. That of type (c) is sometimes referred to as **mixed sampling**. If the universe is divided into "strata" by purposive methods and then a portion of the sample is taken from each "stratum," the sampling is said to be **stratified**.

The application of each of these types may be affected by what is known as **bias**. This is the name given to perturbations which influence the nature of the choice and make it something other than what the experimenter intends it to be. Bias may be due to imperfect instruments, the personal qualities of the observer, defective technique, or other causes. Like experimental error, it is difficult to eliminate entirely, but usually may be reduced to relatively small dimensions by taking proper care.

By an obvious extension of the nomenclature, we talk of a sample obtained by random sampling as a random sample, that obtained by purposive sampling as a purposive sample, and so on.

Random Sampling.

18.18. The reader no doubt already has some intuitive ideas about randomness of choice. We may give a formal definition of random sampling by saying that the selection of an individual from a universe is random when each member of the universe has the same chance of being chosen. Similarly, a sample of n individuals is random when it is chosen in such a way that, when the choice is made, all possible samples of n have an equal chance of being selected.

18.19. The first question arising out of this definition which we have to consider is : How are we to obtain a random sample ?

This question is more difficult than it appears at first sight. It might be thought that any purely haphazard method of selection would give a random sample. For example, if we wished to obtain a random sample of local tradesmen, one way which suggests itself is to take a Trades Directory, open it " at random " and take the first name on which the eye alights, repeating the process until the sample is of the required size. Or again, if we wished to obtain a random sample of wheat growing in a field, it might be thought that a satisfactory method would be to throw a hoop in the air " at random " and select all the plants over which it fell.

18.20. That such methods are apt to be deceptive may be seen from the two examples we have just given. In the first, if we consulted a Trades Directory which had already been used, we should probably find that it opened at some pages more readily than at others ; we should therefore tend to get the more popular tradesmen. Moreover, our eye might tend to be caught by long names or peculiar names. In either case some trades-men would have a greater chance of being chosen than others, and the sample would not be random.

Again, in the second example, our hoop might tend to be caught by the taller ears of wheat, or we might tend unconsciously to throw it towards parts of the field where the wheat looked to be about the average height. These and other factors would destroy the random character of the sampling.

Human Bias.

18.21. Experience has, in fact, shown that the human being is an extremely poor instrument for the conduct of a random selection. Wher-ever there is any scope for personal choice or judgment on the part of the observer, bias is almost certain to creep in. Nor is this a quality which can be removed by conscious effort or training. Nearly every human being has, as part of his psychological make-up, a tendency away from true randomness in his choices.

We may illustrate the unreliability of free choice on the part of even a trained observer by taking an example of height measurements in samples of wheat plants. In the course of certain work at the Rothamsted Experimental Station, sets of eight wheat plants were selected for measure-ment. Six of these shoots were chosen by purely random methods. The other two were chosen " at random " by eye. If, in any set, the eight shoots were ranged in order of magnitude, the two chosen by eye could have any places from one to eight ; and if they, in common with the other six, were really random, they should have occupied these places with equal frequency in a reasonably large number of sets. Table 18.1 shows the resulting frequencies in the ranks one to eight for 116 sets taken on 31st May (before the ears of wheat had formed) and 112 sets taken on 28th June (after the ears had formed).

Fig. 18.1 shows the same results graphically, the dotted line giving the frequencies to be expected if the choice was really random.

The divergence of the actual from the expected results is very striking, and clearly cannot be attributed to fluctuations of sampling. It will be seen that on 31st May, before the ears had formed, the observer was

22

TABLE 18.1.—*Height Measurements of Wheat. Frequencies of Plants Chosen by Eye in Ranks 1-8.* (F. Yates, "Some Examples of Biased Sampling," *Annals of Eugenics*, vol. 6, 1935, p. 202.)

Date.	Observation.	Ascending Order of Magnitude Rank.								Total.	Expectation in Each Class.
		1	2	3	4	5	6	7	8		
May 31	Shoot height	9	7	11	8	11	18	21	31	116	14·5
June 28	Ear height	9	19	27	23	15	10	5	4	112	14
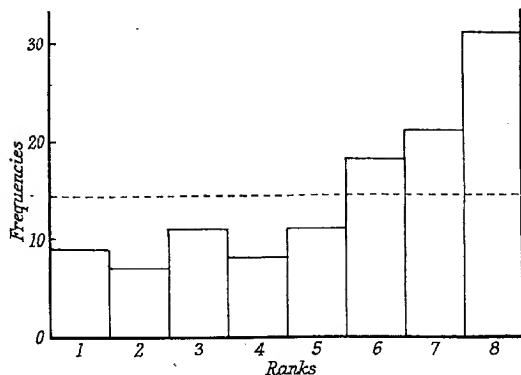
(*a*) Distribution of Shoot Heights (31st May) in Ranks 1-8.
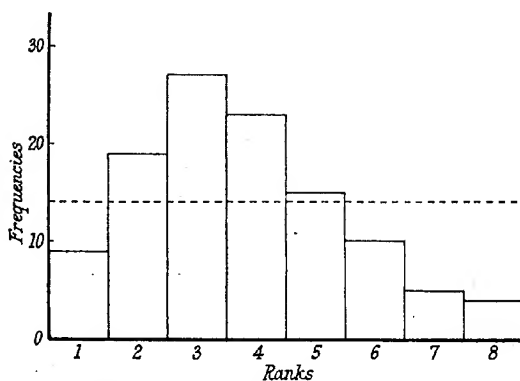
(*b*) Distribution of Ear Heights (28th June) in Ranks 1-8.

FIG. 18.1.—Distribution of Wheat Plants according to Height. (Table 18.1.)

strongly biased towards the taller shoots; whereas in June, after the ears had formed, he was biased strongly towards a central position and avoided short and tall plants.

18.22. Sight is not the only sense which may bias a sampling method. In certain experiments counters of the same shape but of different colours were put into a bag and chosen one at a time, the counter chosen being put back and the bag thoroughly shaken before the next trial. On the face of it this appears to be a purely random method of drawing the counters. Nevertheless, there emerged a persistent bias against counters of one particular colour. After careful investigation the only explanation seemed to be that these particular counters were slightly more greasy than the others, owing to peculiarities of the pigment, and hence slipped through the sampler's fingers.

The student may perform similar experiments for himself. One of the simplest is to ask a friend to recite " at random " one hundred digits, including zero, and then count the number of odd ones. If the numbers are really random, the number of even ones and odd ones should be about equal, but there will frequently be found a bias one way or the other.

18.23. Enough has been said to show that if we are to evolve a satisfactory method of random sampling we must eliminate all personal choice. The method of selection must, therefore, follow some code of procedure which leaves nothing to the observer's idiosyncrasies.

It may sound a little paradoxical to obtain true randomness by following rules of procedure. We are reminded of Bertrand's question: " How can we talk of the laws of chance, which is the negation of all law ? " The ensuing sections will, it is hoped, remove any doubts on this head.

Technique of Random Sampling.

18.24. The methods adopted in any given case to ensure as far as possible that the sampling is random depend to some extent on the size and nature of the universe. Certain modes of procedure which are convenient for small universes are not so for large universes. We shall also see that sampling from a hypothetical universe has a special significance and special difficulties of its own.

18.25. The criterion that every individual should have an equal chance of being chosen may be put in a somewhat different form. If the method of selection is independent of the properties of the sampled universe which it is desired to investigate, there will, so far as those properties are concerned, be no reason why one individual should be chosen rather than another. Hence all values of the properties which occur in the universe will have an equal chance of being chosen. If, therefore, we can produce a mode of procedure which bears no relation to the properties of the parent universe which we are discussing, we may expect that it will give a random sample, so far as those properties are concerned.

18.26. We may now consider a few examples of the kind of procedure to which this rule leads.

Suppose we wish to take a sample of the inhabitants of a street. They are already arranged in houses, and for the sake of simplicity we will take our problem to be that of selecting a number of houses, whose occupants will comprise our sample.

Let us take as our rule of procedure the selection of every tenth house,

starting at some arbitrary point. Unless there are peculiar circumstances, it is presumable that the properties we are investigating, which may, for instance, be income or size of family, are not grouped periodically along the street. The method of selection is then independent of the properties of the universe and the sampling will be random.

If, however, the street were divided into blocks by cross-streets at every tenth house, so that every house in our sample was a corner house, and therefore, possibly, a shop, it is easy to see that the sample is no longer random. Shops occur, in fact, along that street with period ten, and since our method of selection has also that period, the method and the qualities under investigation are no longer independent.

18.27. We might then fall back on a different method. If we take a pack of plain cards, as similar as we can get them, we can make one card correspond to one of the houses by writing on it the number of the house in the street. The pack would then be a kind of miniature of the universe for sampling purposes. We can draw a sample of houses by drawing a sample of cards, and if we shuffle the pack well we have every reason to hope that a random sample will result, for it is hard to imagine any way in which the method of shuffling and drawing could be dependent on the properties of the universe. It is not impossible to make it so, however. For instance, if the ink with which we wrote the numbers on the cards was slightly adhesive, the larger numbers would not be so easy to draw out as the small ones, and we should tend to get houses at one end of the street. If such houses were of the poorer class, our sample for the purpose of investigating income would not be random.

Lottery Sampling.

18.28. The method we have just described, of constructing a miniature universe which is easily handled, is one of the most reliable methods of drawing a random sample. It is the method usually adopted in drawing the winning numbers in sweepstakes and lotteries. In such cases the universe is the aggregate of persons owning tickets in the lottery. To every member of this universe there corresponds a number, the totality of which numbers, written on pieces of paper, comprises the miniature universe. In practice, these pieces are placed in similar containers, usually small metal cylinders, and thrown into a large rotating drum, in which they are thoroughly mixed or " randomised."

18.29. The practical difficulties of constructing the miniature universe and of shuffling it are, however, severe if the parent universe is at all large. The method is, of course, inapplicable on theoretical grounds if the universe is not finite. To save the trouble of work with tickets it is often possible to use numerical methods.

As a rather extreme case, let us consider a method of taking a random sample of the universe of visible stars, which is finite. We will take a star to be defined on the celestial sphere by latitude and longitude, and will ignore difficulties arising from the existence of double stars or unresolved objects. What we want, then, is a set of random pairs of latitudes and longitudes. As a crude method we might take an atlas of the world and choose the figures set out in the index for places arranged alphabetically. But it is easy to see that this method is unsound ; for there will be more names associated with the more populous districts,

and hence the values given in the index will tend to cluster round certain points and avoid others—there will be none in the middle of seas or at the poles, so that the pole star has no chance of being selected.

Let us then take a set of statistical tables and open it haphazardly. We shall be confronted with a page of figures, and if we take, say, the tenth figure in each row we shall probably get a set of digits which are random. Suppose the first ten digits obtained in this way were 7, 0, 4, 7, 9, 6, 8, 2, 9, 1. We might then take our star to be defined by latitude 70° 47·9' and longitude 68° 29·1'. Another page will give us another star, and so on.

Tippett's Numbers.

18.30. The difficulty in applying the method we have just described lies in ensuring that the numbers we obtain are really random. Many tables of figures, such as logarithm tables, may fail to give random digits because there is a relation between the figures in successive rows. To obviate this difficulty certain Tables of Random Sampling Numbers have been constructed by L. H. C. Tippett, by whose name they are known (ref. (605)).

Tippett's numbers consist of 41,600 digits taken from census reports and combined by fours to give 10,400 four-figure numbers. We give here the first forty sets as an illustration of their general appearance :

2952	6641	3992	9792	7979	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2370	7483	3408	2762	3563	1089	6913	7691
0560	5246	1112	6107	6008	8126	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

The reader may wonder how it was ensured that these digits are random. They were chosen haphazard, but the real guarantee of their randomness lies in practical tests. We may say at once that Tippett's numbers have been subjected to numerous investigations which make their randomness for many practical cases highly probable. Their use will be apparent from the following examples :—

Example 18.1.—To take a random sample of 10 from the universe of 8585 men of Table 6.7, page 94.

Here we have 8585 individuals. We will number them from 1 to 8585. The problem of selecting ten men at random is then that of finding ten numbers at random between 1 and 8585. We therefore take a page of Tippett's numbers and select the first ten on the page which are not greater than 8585. Thus, if our page were the one on which appear the numbers we have quoted above, our individuals would be those corresponding to the numbers, reading across,

2952, 6641, 3992, 7979, 5911, 3170, 5624, 4167, 1545, 1396

If we imagine the numbering to be done in order of height, starting with the shortest and ending with the tallest, we see that the first individual falls in the group 66–", the second in the group 69–", and so on. The height-ranges in which the ten individuals fall are, in fact, in inches :

66–, 69–, 67–, 71–, 68–, 66–, 68–, 67–, 65–, 65–

Let us take their heights as being given by the centre points of these ranges, and find their mean.   We have :

$$M - \tfrac{7}{10} = \tfrac{1}{10}(66 + 69 + \ . \ . \ . \ + 65)$$
$$= 67 \cdot 2$$

Hence the mean is 67·6 inches, as against the true value of 67·46 inches in the whole universe.

Example 18.2.—To take a sample of 5 from the distribution of screw lengths of Table 6.3, page 84.

Here we have 206 individuals.   It would clearly be a waste to use only numbers from 0001 to 0206 for the screws and to neglect the rest, and we are able to bring nearly all numbers into play by the following device. We note that 206 goes 48 times into 10,000, with a certain remainder.   In fact, $206 \times 48 = 9888$.   We therefore attach 48 numbers to each screw. Taking them in order, beginning at the shortest, we let the first screw correspond to the numbers 0001 to 0048, the second to 0049 to 0096, the third to 0097 to 0144, and so on, the 206th screw corresponding to the numbers 9841 to 9888.   Numbers above 9888 we leave out of account. Referring to the table, we see that there is one screw in the first category (5 to 6 thousandths short of an inch), four in the second (4 to 5 thousandths short of an inch), and so on.   The numbers corresponding to screws in the different categories will then be 0001–0048, 0049–0240, 0241–0768, and so on ; or, in tabular form,

Difference in Length from 1 inch (thousandths).	Numbers Corresponding.	Difference in Length from 1 inch (thousandths).	Numbers Corresponding.
− 6 to − 5	0001–0048	+ 1 to + 2	5857–7488
− 5 to − 4	0049–0240	+ 2 to + 3	7489–8688
− 4 to − 3	0241–0768	+ 3 to + 4	8689–9456
− 3 to − 2	0769–1824	+ 4 to + 5	9457–9840
− 2 to − 1	1825–3024	+ 5 to + 6	9841–9888
− 1 to 0	3025–4320		
0 to + 1	4321–5856		

We now take five Tippett numbers from the tables.   For instance, we might take the five in the first column of **18.30**, *i.e.* 2952, 4167, 2370, 0560, 2754.   The screws corresponding to these numbers will be 1·5, 0·5, 1·5, 3·5 and 1·5 thousandths short of the inch respectively.

If we had obtained two numbers, say 0001 and 0002 in the first category, we should have been faced with the necessity for a decision on how the sampling was to be regarded, for there is only one screw in this category. If we suppose that a sampled screw is abstracted from the universe, it can only be drawn once ; and hence we should have had to ignore all numbers in the category 0001 to 0048 subsequent to that which first occurs.   If, on the other hand, the screw is replaced, we can draw it as often as we like.

Example 18.3.—In Example 3.5, page 40, we had the following data giving the association between inoculation against cholera and exemption from attack in 818 subjects:—

	Not attacked.	Attacked.	Total.
Inoculated .	276 (0001–3312)	3 (3313–3348)	279
Not inoculated .	473 (3349–9024)	66 (9025–9816)	539
Total .	749	69	818

Let us take a sample of 10 from this universe.

We observe that 818 goes into 10,000 twelve times, with a certain remainder. In fact, $10,000 = 12 \times 818 + 184$. We can therefore attach 12 Tippett numbers to each member of the universe. To the 276 inoculated-not-attacked individuals we attach the numbers 0001 to 3312 (12×276). To the 3 inoculated-attacked individuals we attach the numbers 3313 to 3348 (a range of 36, equal to 3×12). Similarly for the remaining individuals. The Tippett numbers corresponding to the individuals in the four compartments of the table are shown in brackets above.

We then take ten random sampling numbers from the tables, say the first ten, reading across, from the numbers given on page 341. If we had come across a number greater than 9816 we should have ignored it. The first number, 2952, gives us an individual falling in the inoculated-not-attacked class; the second, 6641, gives us a member of the not-inoculated-not-attacked class; and so on. The 10 numbers give the following results :—

	Not attacked.	Attacked.	Total.
Inoculated .	2	0	2
Not inoculated .	6	2	8
Total .	8	2	10

Example 18.4.—Strictly speaking, Tippett's numbers are applicable only to sampling from a finite universe, for we cannot attach a different Tippett number to each member of an infinite aggregate. But, by the following device, we can apply the Tippett tables to draw samples from a continuous (and therefore infinite) universe which is specified by a mathematical equation in such a way as to give us the proportion of the total frequency in given *ranges* of the variate.

In fact, let us draw a sample from a normal universe with unit standard deviation and unit total frequency.

Let us take ranges of 0·1 on each side of the central ordinate. Table 2 of the Appendix will then give us the proportion of the frequency lying in these ranges. As in Example 18.1, we divide up the numbers from 0000 to 9999 in proportion to these frequencies, and this is, in fact, a particularly simple matter. All we have to do, for the positive values of the variate, is to take the figures in the second column (areas) and round them up to four figures. *E.g.* for the first interval 0·0 to 0·1, there will correspond the numbers 5000 to 5398 ; to the interval 0·1 to 0·2, the numbers 5399 to 5793 ; to the interval 0·2 to 0·3, the numbers 5794 to 6179 ; and so on. For the negative values of the variate we have, similarly, for 0·0 to −0·1, the numbers 4601 to 4999; for −0·1 to −0·2, the numbers 4206 to 4600; for −0·2 to −0·3, the numbers 3820 to 4205; and so on, there being as many numbers in any negative range as in the corresponding positive range. Occasionally doubt may arise in assigning a number to a given interval owing to the difficulty of rounding up a figure ending in 5. In practice it is not likely to make any difference which interval we choose; if it threatens to do so, we can take the doubtful number to refer alternately to the two possible intervals.

Having assigned numbers to the ranges, we sample from Tippett's tables in the ordinary way. For instance, a number 5500 will correspond to a member in the range 0·1 to 0·2. If we wish to ascertain the mean of a sample, or some similar function of the variate values, we take the variate value of any individual to be the centre of the interval in which it falls. This is an approximation, but the narrowness of the intervals justifies it in most practical cases.

Further examples will be found in a note by Karl Pearson prefixed to the tables of Tippett numbers themselves. It may be remarked that the tables may be used to give more than 10,400 sets of random four-figure numbers ; we may, for example, construct additional sets by reading the numbers downwards, or taking every other digit diagonally, and so on.

Sampling from Infinite Universes.

18.31. The methods we have just been discussing are appropriate only to those cases in which the universe is finite, so that it was possible to associate with each individual one or more Tippett numbers ; or to universes which, though infinite, can be treated by the method of Example 18.4 owing to their complete specification according to the variate under discussion. The required conditions are met with in much of the material treated in practice, particularly in demographic and economics work ; but in other work the universe may be either infinite or so large as to be infinite for all practical purposes, and a different technique must therefore be used.

Consider, for example, the problem of drawing a random sample from a sack of flour. We clearly cannot number all the particles in the sack, nor could we extract any given particles and examine them. We might, perhaps, reduce this case to that of a finite universe by weighing out the flour into small, say one-ounce, packets and then sampling the packets. This is a kind of mixed sampling. But it is also possible to handle the problem by a special technique, as follows.

First of all, we mix the flour thoroughly. We then divide it into

two halves and select one half. (It does not matter which, but for convenience we may imagine two heaps, one on the right and one on the left, and select left and right alternately.) We then divide the half we have chosen into two further halves, and again select one. The process is continued until the sample has reached a manageable size. We may reasonably suppose that it is random, especially if the flour is well mixed at each stage before being divided into two.

A similar technique may be used for many " continuous " substances, such as milk, grain, cement, etc.

Sampling from Hypothetical Universes.

18.32. The technique for drawing random samples brings out a fundamental difference between existent and hypothetical universes. Taking a simple but typical case, let us draw a sample from the universe of throws of a die.

The methods we have previously used are quite obviously inapplicable here. We cannot construct a card universe, because we do not know the nature of the parent universe. Nor can we put all the possible throws in a heap, and select from it by continued subdivision. In fact, there is only one thing we can do, and that is to throw the die, and take our results as a sample.

What reason have we to suppose that this is a *random* sample ? The answer lies partly in theory and partly in technique. In the first place, we must adapt our method of throwing so that the sampling conditions, so far as we can see, remain constant throughout the experiment. This is a matter of technique, and our methods can, in fact, be tested. But since our universe does not exist for us to examine separately, the only knowledge about it being derived from the sample itself, it will be clear on a little reflection how difficult it is to say that every other possibility in the universe had an equal chance of occurring. We return to this point in **18.35** and **18.36** below.

The Importance of Random Sampling.

18.33. We have already remarked on the importance of being able to gauge the error of an estimate made from a sample. The practical use of the theory of random sampling lies largely in the fact that it allows us to measure objectively, in terms of probability, errors of estimation or the significance of a result obtained from a random sample. The purposive methods to which we refer below do not do this, or at least have not yet been made to do so. The present trend among statisticians is, therefore, on the whole, in favour of the use of random sampling methods except in certain special cases.

18.34. At this point we may bring forward two important considerations.

In the first place, it must not be forgotten that random sampling *may* produce the most unrandom-looking results. For instance, we usually regard a hand of cards at bridge as a random sample from the universe of 52 which comprise the pack ; but it is not unknown for a hand of 13 spades to be dealt. The fact that the sample looks purposive, therefore, *proves* nothing. But it does provide a basis for strong presumptions. How strong those presumptions may be the student may judge for himself

by imagining what he would think of a card party at which he got 13 spades twice in succession!

Secondly, we can never be absolutely certain that a method of sampling is random. There are doubts on *a priori* grounds because for any given method there are always *conceivable* sources of bias, and we can never rule out entirely the possibility that some of these sources are present. The utmost we can do is to make their presence extremely unlikely by taking great care with the experiment.

18.35. We can, however, apply tests to judge the randomness of a sampling method. If we draw a single sample from a known universe, the result will tell us nothing about the method adopted ; but if we take a large number of samples they should, if the sampling is random, be distributed in a certain way, and for some universes we can calculate mathematically what that way ought to be. If, therefore, we apply our sampling method to such a parent universe and find the results widely divergent from expectation, we have every reason to suspect our sampling technique. *Per contra,* if the results and expectation are in accord, there is good ground for reliance on the sampling.

18.36. Tests of this kind presuppose that we know the form of the parent universe. In sampling from a hypothetical universe we do not know this, and are forced to estimate it from the sample. Clearly, we cannot use this estimate to criticise the method by which the sample was obtained without some closer inquiry.

Similar problems may arise for existent universes when we do not know the nature of the parent universe but have to estimate some or all of its characteristics from the data of the sample. In such cases it is extremely difficult to be completely satisfied that the sampling is random. Frequently the best we can do is to use a method which has been found satisfactory for other universes and hope, in the absence of any indication to the contrary, that it will also be satisfactory for the present universe.

Purposive Sampling.

18.37. We have already pointed out the dangers of introducing bias if the observer gives rein to his inclinations in choosing a sample, and have stressed the fact that in general there does not exist a method of assessing the degree of accuracy of an estimate made from a purposive sample. In spite of these handicaps, however, there are cases where purposive selection is a useful method. In this book we shall not consider it in any great detail, because the reliance placed upon it depends largely on the circumstances of the case, remains to a great extent a matter of personal opinion, and is not capable of being discussed by elementary methods. Nevertheless, our brief survey would be incomplete without some reference to it.

18.38. Let us first of all consider the case of an observer who wishes to take a sample of two or three turnips from a cart-load. A *random* sample might give us several very large or very small turnips, though it is unlikely to do so. But if we allow the observer to run his eye over the whole load and then choose, he is most likely to take what he regards as average turnips—*i.e.* average in size, weight, shape, and whatever other quality may be in his mind.

It may be claimed, with some plausibility, that this purposive method

is more likely to give us a sample which is *typical* or *representative* of the universe than a random method. The random sample may vary widely from the average, whereas the purposive sample does not. This gives the latter an advantage as a rule ; but it may be pointed out—

(*a*) That as the sample becomes larger the random sample becomes more and more representative of the parent, whereas, owing to bias, the purposive sample in general does not.

(*b*) That in many cases the object of the sample is to give us information about the whole of the universe ; the purposive sample might tell us more about the mean weight of the turnips, but would probably give a worse idea of the variance of the weights because the observer has deliberately chosen values near the mean.

18.39. If we had to choose between pure random sampling and purposive sampling, our choice would probably be determined by balancing the uncertainties of the former, which are mainly due to fluctuations of chance, and the uncertainties of the latter, which are mainly due to bias. In practice, however, it is often possible to combine the two methods in **stratified sampling** and gain some of the advantages of each while minimising their disadvantages.

The essentials of this process lie in dividing the parent population into strata and taking a random sample from each stratum. For instance, if we are taking a sample of earned incomes, we might first group individuals into classes " earning up to £500 per annum," " earning from £500 to £1000 per annum," and so on, and then choose a random sample from each class. Or, if we wanted a sample of farms in Great Britain, we might first classify them roughly as " devoted mainly to arable crops," " devoted mainly to milk production," " devoted mainly to vegetable growing," etc., and again take a random sample from each group.

18.40. Finally, we may also sample a universe by first of all arranging its individuals in groups. This amounts to taking a different sampling unit. For instance, in sampling the population of Great Britain we might, as a matter of convenience, take streets or local government districts instead of individual human beings as our unit. We have already had an instance of this type when we suggested as one way of sampling a sack of flour that it might be weighed out first into one-ounce packets. The process is obviously more convenient when this grouping has been done for us, *e.g.*, in census returns.

18.41. Each branch of science and industry presents its own sampling problems, and it would be difficult to expand the foregoing discussion so as to include the detailed requirements of the worker in every sphere. We will conclude this chapter with an example of the way in which all the methods we have described may be pressed into service in order to give a sample which is as representative as practical limitations will allow.

It is the practice in England for manufacturers of sugar from sugar beet to pay the growers according to the sugar content of their product. The beet, which is not unlike a parsnip, is delivered to the factory in lots of at least several tons with a certain amount of waste material, such as earth, adhering to it. The problem is, then, (*a*) to find the net weight of the beet when cleaned and ready for the slicing process, which is the first stage in

the extraction of the sugar, and (b) to ascertain the sugar content. The method of procedure is as follows :—

The gross weight of the load of beet usually is first obtained by weighing the lorry which contains it when full, and when empty. From the middle of the load of beet is then abstracted about 28 pounds, which is carefully weighed, and then cleaned and weighed again. The difference in the weights gives the " tare," that is to say, the proportion of waste matter, and a proportional amount is deducted from the whole load to give the net weight of beet. This process is equivalent to taking a random sample and assuming that the value of the " tare " in the sample is the value in the whole universe.

The sample of washed beet is then laid out on a table and arranged with the roots in order of size. From this sample a smaller sample is taken by choosing a beet every so often. This is a process of pure purposive selection.

The reduced sample is still inconveniently large, so it is reduced by taking a slice from each beet. It is known that the sugar in the root is not distributed homogeneously (although it is roughly symmetrical about the axis of the root), so trained men are employed to slice one section with a rasp, the section being that which would be obtained by cutting the root from the thick end to the tapered end into two symmetrical halves and then repeating the process one or more times. This selection again is purposive in so far as the shape of the section is based on knowledge of the distribution of the sugar, but random in so far as it is a matter of chance what is the longitude of the particular slice chosen.

When each beet has been treated in this way there is given a heap of pulp which may be analysed. The heap is, however, as a rule still too large. It is therefore well mixed and divided into four heaps. Two heaps are thrown away, one is reduced to 26 grammes and analysed by the factory and one, similarly reduced, is analysed by the grower's representative. This last method of selection is a random method adapted for a universe which cannot readily be enumerated.

The final sample therefore appears as the result of four successive sampling methods, two of which are random, one purposive, and one a mixture of purposive and random.

SUMMARY.

1. Sampling may be random, purposive or mixed.

2. Random sampling owes its importance to the fact that we can assess the results obtained from it in terms of probability.

3. The presence of an element of choice on the part of the observer introduces the danger of bias, and should not be permitted where it can be avoided.

4. Random samples may conveniently be drawn by the use of card universes or of Tippett's numbers.

5. The sampling technique adopted in any given case will depend largely on the circumstances of that case and the resources of the observer. At the present time the reliability of estimates made from samples is partly a matter of individual opinion founded on intuitive ideas, unless the sampling methods are random.

EXERCISES.

18.1. Draw a random sample of 20 from the universe of men of the last column of Exercise 6.6 (inhabitants of the United Kingdom classified according to weight). Find the mean of the sample and compare it with the mean of the universe.

18.2. Deal yourself a hand of 13 cards from an ordinary pack of 52 playing cards and count the number of court cards. Use your result to estimate the number of court cards in the whole pack.

Repeat the experiment ten times, taking a new deal each time, and compare the mean of your results with the true value, 12.

18.3. Suggest a method for obtaining a random sample of words from the English language by the use of Tippett's numbers and a dictionary.

18.4. Draw a sample of 30 from the universe of the last column of Table 6.7, and find the standard deviation. Compare your result with the standard deviation of the universe.

18.5. Suggest a possible source of bias in the following:—

(a) A barrel of apples is sampled by taking a handful from the top.

(b) A mixture of sand and sawdust is sampled by scooping up a quantity from the bottom.

(c) A set of digits is taken by opening a Telephone Directory at random and choosing the telephone numbers in the order in which they appear on the page.

(d) Readers of a newspaper are sampled by printing in it an invitation to them to send up their observations on some topical event.

(e) Investigators into the size of families in a town conduct a house-to-house inquiry (1) in the morning, (2) in the afternoon, ignoring those houses at which there is no reply.

18.6. Draw 100 samples of 10 from a normal universe by means of Tippett's numbers, and form the frequency-distribution of their means.

18.7. In the data obtained in Exercise 18.6, form the frequency-distribution of the root-mean-square deviations of the samples about the mean of the parent universe.

18.8. Draw 100 samples of 10 from the Poisson universe of 10.47, page 191, and form the frequency-distribution of their means.

18.9. Draw 500 samples of 4 from the universe of Australian marriages of Table 6.8, page 96, and form the frequency-distribution of their range.

18.10. Draw a sample of 50 from the universe of Table 11.4, page 200 (4912 dairy cows), and find the correlation in the sample between age in years and yield of milk per week. Compare your result with the correlation in the universe.

CHAPTER 19.

THE SAMPLING OF ATTRIBUTES—LARGE SAMPLES.

The Problem.

19.1. In dealing with the theory of sampling we shall find it convenient to preserve the formal distinction between attributes and variables which we drew earlier in this book. The theory of the sampling of attributes is in many respects simpler than that of variables, and in this chapter we shall confine ourselves to it. We shall begin by considering a type of sampling which we shall call **simple**, involving certain limitations on the generality of the problem, and shall then proceed to examine the removal of these limitations in order to deal with the general case.

19.2. The sampling of attributes may be regarded as the drawing of samples from a universe containing A's and not-A's. The number of A's in each sample, or the proportion of A's, will form part of the data provided by the samples.

We shall find it convenient to adopt the nomenclature of **10.3** and to speak of the drawing of an individual on sampling as an " event." The appearance of the attribute A may be called a " success " and the non-appearance a " failure." Thus, in sampling a human population for the proportions of the two sexes, we might say of a sample of 100, 45 of which were male, that the sample consisted of 100 events, 45 of which were successes and 55 failures. (It might, of course, be more convenient—and would certainly be more courteous—to reverse the names and call the occurrence of a female a " success " and of a male a " failure.")

Simple Sampling.

19.3. By **simple** sampling we mean random sampling in which each event has the same chance p of success, and in which the chances of success of different events are independent, whether previous trials have been made or not. These conditions hold good, for instance, in the throwing of a die or the tossing of a coin ; the chance of getting heads with a coin is not affected by what was obtained on the previous trials, and remains constant no matter how many trials are made, provided, of course, that the coin does not begin to wear or is not falsely manipulated by the experimenter.

Simple sampling is a particular form of random sampling, as we have defined it in the previous chapter. Suppose, for example, we take a sample of two from a universe consisting of 6 men and 4 women under random sampling conditions, *i.e.* so that at each of the two events which constitute the sample every member of the universe has an equal chance of being chosen. If, at the first trial, we draw a man, the chance of doing so being $\frac{6}{10}$, there will be 5 men and 4 women left in the universe, and the chance of obtaining a man on the second trial will be $\frac{5}{9}$. This is not the same as the chance on the first trial, and hence the sampling is not simple, though it is random.

Mean and Standard Deviation in Simple Sampling of Attributes.

19.4. Suppose now that we take N samples with n events in each. The chance of success of each event is p and of its failure $q = 1 - p$. As in **10.6**, the frequencies of samples with 0, 1, 2, . . . successes are the terms in the series $N(q + p)^n$, *i.e.*

$$N\left\{q^n + nq^{n-1}p + \frac{n(n-1)}{2}q^{n-2}p^2 + \ldots + nqp^{n-1} + p^n\right\}$$

As in **10.9**, this distribution has mean M given by

$$M = np$$

and standard deviation (**10.10**)

$$\sigma = \sqrt{npq} \qquad . \qquad . \qquad . \qquad . \quad (19.1)$$

19.5. In lieu of recording the number of successes in each sample we might have recorded the proportion of successes, that is, $\frac{1}{n}$-th of the number in each sample. As this would amount to dividing all figures of the record by n, the mean proportion of successes must be p, and the standard deviation of the proportion of successes is given by

$$s = \sqrt{\frac{pq}{n}} \qquad . \qquad . \qquad . \qquad . \quad (19.2)$$

Equations (19.1) and (19.2) are of fundamental importance.

Example 19.1.—The following results, due to Weldon, are of interest. Weldon threw 12 dice 4096 times, a throw of 4, 5 or 6 being called a success. We have, then, 4096 samples of 12 from the universe consisting of all possible throws of the dice.

If the dice are all true, the chance of success is $\frac{1}{2}$. Hence, the theoretical mean $M = 6$; theoretical value of the standard deviation $\sigma = \sqrt{0 \cdot 5 \times 0 \cdot 5 \times 12} = 1 \cdot 732$.

The following was the frequency-distribution observed : –

Successes.	Frequency.		Successes.	Frequency.
0	—		7	847
1	7		8	536
2	60		9	257
3	198		10	71
4	430		11	11
5	731		12	—
6	948			
			Total	4096

Mean $M = 6 \cdot 139$, standard deviation $\sigma = 1 \cdot 712$. The proportion of successes is $6 \cdot 139/12 = 0 \cdot 512$ instead of $0 \cdot 5$.

Example 19.2.—(G. U. Yule.) The following may be taken as an illustration based on a smaller number of observations : Three dice were thrown 648 times, and the numbers of 5's or 6's noted at each throw. $p = 1/3$, $q = 2/3$; theoretical mean 1; standard deviation 0·816.

Frequency-distribution observed :

Successes.	Frequency.
0	179
1	298
2	141
3	30
Total	648

$M = 1\cdot034$, $\sigma = 0\cdot823$. Actual proportion of successes $0\cdot345$.

19.6. The value pn is sometimes called the " expected " value of the number of successes in the sample. It is not only the mean value of all samples, but is the most probable value and is also representative, *i.e.* it bears the same ratio p to the number in the sample as the number of individuals with attribute A in the universe bears to the total number in the universe. The divergences of the number of successes from the expected value in any given random sample give rise to what we have hitherto called fluctuations of random sampling. They are to be regarded as deviations due to the nature of the sampling process, and not indicative of any real properties of the universe itself.

19.7. Equations (19.1) and (19.2) enable us to deal with the question which has arisen several times in earlier chapters of this book, namely, when can we say that observed deviations from the expected values in a sample of attributes are due to some real effect and are not merely attributable to sampling fluctuations ?

The binomial distribution, to which samples classified according to the frequencies of an attribute give rise, is a single-humped type which approximates very closely to the normal for large values of n, the number in the sample. It follows that the great majority of its members lie within a range $\pm 3\sigma$ on each side of the mean, *i.e.* of $\pm 3\sqrt{npq}$ on each side of the value np. If the distribution is exactly normal, $0\cdot9973$ of the curve lies within this range **(10.29)**. We can therefore say that if a particular sample gives a value of p outside this range, the deviation from the expected value is most unlikely to have arisen from fluctuations of simple sampling. If n is large, the chances are about 3 in a thousand that it arose in that way.

It must be emphasised that the free use of the 3σ rule is justified only if n is large.

Example 19.3.—In the experiments of Example 19.1, 25,145 throws of a 4, 5 or 6 were made out of 49,152 throws altogether. The chance of throwing one of these numbers is $\frac{1}{2}$, and hence the expected value is 24,576. The observed number was thus 569 in excess of this. Can the deviation from the expected value be due to fluctuations of simple sampling ?

The standard deviation of simple sampling is

$$\sigma = \sqrt{npq} = \sqrt{\tfrac{1}{2} \times \tfrac{1}{2} \times 49152}$$
$$= 110\cdot9$$

The deviation observed is 5·13 times this quantity, and it is therefore most improbable that it arose as a sampling fluctuation.   We must therefore seek some other explanation of the deviation, and it seems reasonable to suspect that the dice were slightly biased.

The problem might, of course, have been attacked equally well from the standpoint of *proportion* instead of the actual numbers of successes. This proportion is 0·5116 instead of the expected 0·5000, the difference in excess being 0·0116.   The standard deviation of the proportion is

$$s = \sqrt{\frac{1}{2} \times \frac{1}{2} \times \frac{1}{49152}} = 0.00226$$

and the difference observed is 5·13 times this, which is the same ratio as before, as of course it must be.

Example 19.4.—(Data from the *Second Report of the Evolution Committee of the Royal Society*, 1905, p. 72.)

Certain crosses of the pea, *Pisum sativum*, gave 5321 yellow and 1804 green seeds.   The expectation is 25 per cent. of green seeds on a Mendelian hypothesis.   Can the divergences from the expected values have arisen from fluctuations of simple sampling ?

The numerical difference from the expected result is 23.   The standard deviation of simple sampling is

$$\sigma = \sqrt{0.25 \times 0.75 \times 7125} = 36.6$$

The divergence from theory is only about 0·6 of this, and hence may very well have arisen from fluctuations of simple sampling.

Standard Error.

19.8.   We shall very frequently have to use the standard deviation of sampling, and it is convenient to have a shorter name for this quantity. We shall call it the **standard error.**   The use of the word error is justified in this connection by the fact that we usually regard the expected value as the true value, and divergences from it as errors of estimation due to sampling effects ; but the student should not attach too much significance to the particular term " error."

In most of our work the term " standard error " will be applied to the standard deviation of *simple* sampling ; but it has a rather wider meaning, embracing this one, which we shall discuss in considering the sampling of variables (**20.22,** *cf.* also **19.31**).

We may, then, summarise the foregoing in the statement that frequencies differing from the expected frequency by more than 3 times the standard error are almost certainly not due to fluctuations of sampling. They point to some departure of the sampling from simplicity, which may in turn point either to some flaw in the sampling technique or to causal effects in the universe itself.

Probable Error.

19.9.   Instead of the standard error, some authorities have used a quantity called the *probable error*, which is 0·67449 times the standard error.   This practice arose from the fact that in the normal curve the

quartiles are distant 0.67449σ from the mean, so that the probability that a deviation is in excess of the probable error is $\frac{1}{2}$, and is equal to the probability of a deviation being less than the probable error. The rule that the observed deviation should not be greater than 3 times the standard error is then approximately equivalent to a rule that it should not exceed 4·5 times the probable error.

The use of the probable error is declining, and we recommend the student to eschew it.

19.10. In Examples 19.1 to 19.4 we dealt with cases where p, the probability of success, was known *a priori*. In many cases it is not known, and further consideration is necessary before we can apply equations (19.1) and (19.2) to such cases.

To fix the ideas, let us suppose that we have a simple sample of 1000 individuals from the inhabitants of Great Britain, and find that 36 per cent. of them have blue eyes and the remainder have eyes of some other colour. What can we infer about the proportion of blue-eyed individuals in the whole population ?

In this instance we do not know the proportion p of blue-eyed individuals in the population. We do know that the standard error is $\sqrt{1000pq}$. Now, whatever p and q are, pq cannot exceed $\frac{1}{4}$, and hence the standard error cannot exceed $\frac{1}{4}\sqrt{1000}$, or 16. Hence, whatever p is, a simple sample should give a number of successes within 3 times this, or 48, of the expected frequency pn. This is 4·8 per cent. of the sample, and we thus may say that the proportion of blue-eyed people in the whole population is 36 ± 4.8 per cent., *i.e.* that it lies between 31·2 and 40·8 per cent.

19.11. We may, however, make a rather better estimate. We have seen that the standard error is small compared with the expected value, and hence with the observed value. If, therefore, in calculating the standard error we take the observed values of p and q *in the sample* instead of the unknown true values of p and q, we shall not involve ourselves in very great error.

Thus, taking p to be 0·36, $q = 0.64$,

$$\sigma = \sqrt{npq} = \sqrt{0.36 \times 0.64 \times 1000}$$
$$= 15.18$$

Hence, $3\sigma = 45.5$ approximately, and the limits are now 36 ± 4.6 or 31·4 and 40·6—slightly narrower than those previously obtained.

19.12. In this example we have taken the proportion of successes in the sample to be an estimate of the proportion of successes in the universe, and have set limits to the range within which the true proportion probably lies. There are other reasons, of an advanced theoretical character which we shall not specify, for taking p in the sample as an estimate of p in the universe, but the student will probably concede that it is the most reasonable thing to do in the circumstances. We must, however, look a little more closely into the assumption that this estimate may be used in calculating the standard error.

19.13. The assumption is a justifiable one if n is large and neither p nor q is small. For in such a case, the standard error of the proportion p is $\sqrt{\dfrac{pq}{n}}$, and this is small compared with p unless p itself is small.

If, then, the standard error of p is small, the value of p estimated from the sample must be close to the real value, and we shall not introduce any serious error by taking the estimated value in evaluating the formula $\sqrt{\dfrac{pq}{n}}$.

19.14. Precisely how large n must be for this approximation to be valid it is not easy to say. Samples of 1000 are almost certainly large enough, and we may often apply the foregoing procedure with considerable confidence to much smaller samples, say of 100. For samples below that figure it is as well to examine carefully the circumstances of any given case and to proceed with caution.

We shall have more to say on this matter when we consider the sampling of variables **(20.17** and **20.18)**.

For the remainder of this chapter we shall assume that our samples are " large," that is to say, that the approximations involved in our assumptions as to the estimate of p are valid.

Example 19.5.—A sample of 900 days is taken from meteorological records of a certain district, and 100 of them are found to be foggy. What are the probable limits to the percentage of foggy days in the district ?

Anticipating somewhat our discussion of simple sampling, we will assume that the conditions of this problem give a simple sample.

Hence,

$$p = \tfrac{1}{9}, \qquad q = \tfrac{8}{9}$$

Standard error of the proportion of foggy days

$$= \sqrt{\frac{pq}{n}} = \sqrt{\frac{1}{9} \times \frac{8}{9} \times \frac{1}{900}}$$

$$= 0 \cdot 0105$$

$$= 1 \cdot 05 \text{ per cent.}$$

Hence, taking $\tfrac{1}{9}$ to be the estimate of the number of foggy days, we have that the limits are $11 \cdot 11$ per cent. $\pm\ 3 \cdot 15$ per cent., *i.e.* 8 per cent. and $14 \cdot 25$ per cent. approximately.

Example 19.6.—A biased penny is tossed 100 times and comes down heads 70 times. What are the probable limits to the probability of getting a head in a single trial ?

We require to know the limits of p. If we assume that 100 is a large sample, we have :

$$\sqrt{\frac{pq}{n}} = \sqrt{\frac{1}{100} \times \frac{7}{10} \times \frac{3}{10}} = 0 \cdot 0458$$

The limits are therefore $0 \cdot 70 \pm (3 \times 0 \cdot 0458)$

$$= 0 \cdot 70 \pm 0 \cdot 1374$$

$$= 0 \cdot 56 \quad \text{and} \quad 0 \cdot 84 \text{ approximately}$$

If we feel any doubt as to the validity of using estimates of p and q from a sample of 100 in calculating the standard error, we may proceed as follows :—

The standard error of p cannot exceed $\sqrt{\frac{1}{100} \times \frac{1}{2} \times \frac{1}{2}}$, *i.e.* 0·05. Hence the value of p lies almost certainly within the limits 0·70 ± 0·15, *i.e.* 0·55 and 0·85.

$$\text{If } p = 0.55, \qquad \sqrt{\frac{pq}{n}} = 0.04975$$

$$\text{If } p = 0.85, \qquad \sqrt{\frac{pq}{n}} = 0.03571$$

For intermediate values of p, $\sqrt{\frac{pq}{n}}$ lies between these limits. Hence the maximum value of the standard error is 0·04975, and p lies between the limits 0·70 ± 0·14925, *i.e.*

$$0.55075 \quad \text{and} \quad 0.84925$$

It will be seen that these limits are nearly equal to those obtained on the assumption that $p = q = \frac{1}{2}$, and are not very different from those we got by assuming $p = 0.70$. There would, however, be an appreciable difference if p had been small, say 0·10.

19.15. If one of the two proportions p and q becomes very small, equation (19.1) may be put into an approximate form that is very useful. Suppose p to be the proportion that becomes very small, so that we may neglect p^2 compared with p; then

$$pq = p - p^2 = p \text{ approximately}$$

and consequently we have approximately:

$$\sigma = \sqrt{np} = \sqrt{M} \qquad . \qquad . \qquad . \qquad (19.3)$$

That is to say, *if the proportion of successes be small, the standard deviation of the number of successes is the square root of the mean number of successes.* Hence we can find the standard error even though p be unknown, provided only we know that it is small.

This is, in fact, the case when the binomial becomes the Poisson series (**10.40**). For such distributions the rule that a range of 6σ includes the great majority of the observations remains valid, as may be seen from the diagram on page 190, but the limits assigned to the standard error of the mean M may be too wide on the left of the mean. For example, if $M = 1$, $\sigma = 1$, and a range of 3 units to the left of the mean carries us to a value of -2, whereas there can be no part of the frequency with negative values of the variate.

19.16. It will be noticed that the standard error depends only on the value of p and the size of the sample, and that therefore the range within which p probably lies is independent of the size of the universe. This appears a little paradoxical, because one might expect that a sample which was, say, 20 per cent. of the universe would enable closer limits to be set than one which was 10 per cent. of the universe.

The explanation is to be found in the nature of simple sampling itself. We shall see below that the conditions under which simple sampling arises in practice are such that either the universe is actually or practically infinite, or each member drawn for a sample is put back in the universe

before the next is drawn. In either case the universe is inexhaustible, and no sample is any nearer to including all its members than another sample. It is, therefore, not surprising to find that the size of the universe does not appear in the formula for the standard error.

19.17. A further notable fact is that the standard error of p varies inversely as the square root of n, and not inversely as n itself. Thus, as n becomes larger the standard error becomes smaller, which is what we should expect, but the standard error decreases proportionately to the square root of n. For instance, if a sample of 100 gives us a standard error of 10 per cent., it will take a sample of 400 to halve that error, and a sample 100 times as large, *i.e.* 10,000, to reduce the error to one-tenth or one per cent.

Precision.

19.18. The standard error may fairly be taken to measure the unreliability of an estimate of p; the greater the standard error, the greater the fluctuations of the observed proportion, although the true proportion is the same throughout. The reciprocal of the standard error ($1/s$), on the other hand, or some convenient multiple of the reciprocal—*cf.* 8.15 and 10.32—may be regarded as a measure of *reliability*, or, as it is sometimes termed, *precision*, and consequently *the reliability or precision of an observed proportion varies as the square root of the number of observations on which it is based.*

The Limitations of Simple Sampling.

19.19. In order to realise the limitations on the use of the formulæ of equations (19.1) and (19.2), it is necessary to consider what are the conditions which will give rise to simple sampling in practice. Supposing, for example, that we observe among groups of 1000 persons, at different times or in different localities, the various percentages of individuals possessing certain characteristics—dark hair, or blindness, or insanity, and so forth. Under what conditions should we expect the observed percentages to obey the law of sampling that we have found, and show a standard deviation given by equation (19.2) ?

19.20. In the first place, the condition that p, the probability of drawing an individual with attribute A on random sampling, remains constant, and in particular is the same for all samples, means that the proportion of individuals with attribute A in the universe must remain constant at the drawing of each sample. Consequently, if formula (19.2) is to hold good in our practical case of sampling there must not be a difference in any essential respect—*i.e.* in any character that can affect the proportion observed—between the localities from which the samples are drawn, nor, if the samples have been made at different epochs, must any essential change have taken place during the period over which the observations are spread. Where the causation of the character observed is more or less unknown, it may, of course, be difficult or impossible to say what differences or changes are to be regarded as essential, but where we have more knowledge the condition laid down enables us to exclude certain cases at once from the possible applications of formula (19.1) or (19.2). Thus it is obvious that the theory of simple sampling cannot apply to the variations of the death-rate in localities with populations

of different age and sex composition, or to death-rates in a mixture of healthy and unhealthy districts, or to death-rates in successive years during a period of continuously improving sanitation. In all such cases variations due to definite causes are superposed on the fluctuations of sampling.

19.21. Secondly, the proportion of individuals with attribute A must remain constant for the drawing of each individual member of the sample. This is again a very marked limitation. To revert to the case of death-rates, formulæ (19.1) and (19.2) would not apply to the numbers of persons dying in a series of samples of 1000 persons, even if these samples were all of the same age and sex composition, and living under the same sanitary conditions, unless, further, each sample only contained persons of one sex and one age. For if each sample included persons of both sexes and different ages, the condition would be broken, the chance of death during a given period not being the same for the two sexes, nor for the young and the old. The groups would not be homogeneous in the sense required by the conditions from which our formulæ have been deduced.

19.22. We pointed out in **19.3** that sampling from a finite universe is not simple owing to the fact that the abstraction of an individual alters the chance of success at the next trial. In practice there are three important cases in which the condition for the constancy of p is satisfied :

(*a*) If the individuals are replaced at each drawing before the next drawing is made ; for in this case the constitution of the universe is the same at each trial, and hence the chance of success must also be the same.

(*b*) If the universe is infinite ; for in this case the withdrawal of a finite number of members does not affect the proportion of individuals in the universe possessing the attribute in question.

(*c*) If the universe is very large, p may be taken to be constant without sensible error, provided that the sample is not also large. This is a very important case, and justifies the application of the theory of simple sampling to many practical data.

Suppose, for instance, we are sampling the population of the United Kingdom for sex ratio, and decide to take a sample of 1000. Suppose again, for the purposes of illustration, that the whole population consists of 23 million women and 22 million men. The chance of getting a man at the first trial will then be $\frac{22,000,000}{45,000,000}$. If we succeed in getting a man, the chance of doing so at the second trial will be $\frac{21,999,999}{44,999,999}$. Even if we draw 999 men the chance of success at the thousandth trial would be $\frac{21,999,001}{44,999,001}$. All these chances, to a close approximation, are equal, and we can assume them to be so without fear of appreciable error. The case would, of course, have stood differently if our sample had numbered several millions.

19.23. A third condition for simple sampling was explicitly stated in our definition in **19.3**. The individual events must be completely independent of one another, like the throws of a die, or sensibly so, like the drawing of balls from a bag containing a number of balls which is large

compared with the number drawn. Reverting to the illustration of a death-rate, our formulæ would not apply even if the sample populations were composed of persons of one age and one sex, if we were dealing, for example, with deaths from an infectious or contagious disease. For if one person in a certain sample has contracted the disease in question, he has increased the possibility of others doing so, and hence of dying from the disease. The same thing holds good for certain classes of deaths from accident, *e.g.* railway accidents due to derailment, and explosions in mines : if such an accident is fatal to one person it is probably fatal to others also, and consequently the annual returns show large and more or less erratic variations.

19.24. It is evident that these conditions very much limit the field of practical cases of an economic or sociological character to which formulæ (19.1) and (19.2) can apply without considerable modification. The formulæ appear, however, to hold to a high degree of approximation in certain biological cases, notably in the proportions of offspring of different types obtained on crossing hybrids, and, with some limitations, to the proportions of the two sexes at birth. It is possible, accordingly, that in these cases all the necessary conditions are fulfilled, but this is not a necessary inference from the mere applicability of the formulæ. In the case of the sex ratio at birth it seems doubtful whether the rule applies to the frequency of the sexes in individual families of given numbers, but it does apply fairly closely to the sex ratios of births in different localities, and still more closely to the ratios in one locality during successive periods. That is to say, if we note the number of males in a series of groups of n births each, the standard deviation of that number is approximately $\sqrt{npq}$, where p is the chance of a male birth ; or, otherwise, $\sqrt{pq/n}$ is the standard deviation of the proportion of male births.

Applications of Simple Sampling.

19.25. We have already shown in examples how the theory of simple sampling can be used to gauge the precision of an estimate of the proportion of individuals in a universe which possess an attribute A, and to set limits outside which that proportion probably does not lie. We now turn to further applications of the theory in the checking and control of the interpretation of statistical results.

19.26. *Case 1.* —Given the expected frequency in a sample and the observed frequency of successes, it is desired to know whether the deviation of the second from the first can have arisen from fluctuations of simple sampling.

This is a case which we have discussed in Examples 19.3 and 19.4. From the expected frequency we can calculate the standard error, and if the deviation is more than 3 times this quantity it almost certainly did not arise from fluctuations of random sampling.

19.27. One caution is necessary here. If the deviation is less than 3 times the standard error, it does not follow that the expected frequency divided by the number in the sample is really the proportion of individuals possessing the attribute A in the universe. In other words, if the expected value is derived from some hypothesis, such as the Mendelian hypothesis in the case of Example 19.4, the fact that the deviation lies within the limits of 3 times the standard error does not prove the hypothesis correct. It

only indicates that experiment and hypothesis are not in disagreement. Furthermore, if the deviation lay without those limits, the hypothesis would not necessarily be disproved, for the fault might lie with the randomness of the sampling.

19.28. *Case 2.*—Two samples from distinct materials or different universes give proportions of A's p_1 and p_2, the numbers of observations in the samples being n_1 and n_2 respectively. (*a*) Can the difference between the two proportions have arisen merely as a fluctuation of simple sampling, the two universes being really similar as regards the proportion of A's therein ? (*b*) If the difference indicated were a real one, might it vanish, owing to fluctuations of sampling, in other samples taken in precisely the same way ? This case corresponds to the testing of an association which is indicated by a comparison of the proportion of A's amongst B's and β's.

(*a*) We have no theoretical expectation in this case as to the proportion of A's in the universe from which either sample has been taken.

Let us find, however, whether the observed difference between p_1 and p_2 may not have arisen solely as a fluctuation of simple sampling, the proportion of A's being really the same in both cases, and given, let us say, by the (weighted) mean proportion in our two samples together, *i.e.* by

$$p_0 = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

(the best guide that we have).

Let ϵ_1, ϵ_2 be the standard errors in the two samples, then

$$\epsilon_1{}^2 = p_0 q_0 / n_1, \qquad \epsilon_2{}^2 = p_0 q_0 / n_2$$

If the samples are simple samples in the sense of the previous work, then the mean difference between p_1 and p_2 will be zero, and the standard error of the difference ϵ_{12}, the samples being independent, will be given by

$$\epsilon_{12}^2 = p_0 q_0 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \qquad . \qquad . \qquad . \qquad (19.4)$$

If the observed difference is less than some three times ϵ_{12}, it may have arisen as a fluctuation of simple sampling only.

(*b*) If, on the other hand, the proportions of A's are not the same in the material from which the two samples are drawn, but p_1 and p_2 are the true values of the proportions, the standard errors of sampling in the two cases are

$$\epsilon_1{}^2 = p_1 q_1 / n_1, \qquad \epsilon_2{}^2 = p_2 q_2 / n_2$$

and consequently

$$\epsilon_{12}^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} \qquad . \qquad . \qquad . \qquad (19.5)$$

If the difference between p_1 and p_2 does not exceed some three times this value of ϵ_{12}, it may be obliterated by an error of simple sampling on taking fresh samples in the same way from the same material.

The student will note that in arriving at these results we have assumed that the unknown values p_0, p_1, p_2 are given to a sufficient degree of approximation by estimates from the samples. This, as we have seen, is justified if n be large.

Example 19.7.—(Data from J. Gray, "Memoir on the Pigmentation Survey of Scotland," *Jour. of the Royal Anthropological Institute*, vol. 37, 1907.) The following are extracted from the tables relating to hair-colour of girls at Edinburgh and Glasgow :—

	Of Medium Hair-colour.	Total observed.	Per cent. Medium.
Edinburgh . .	4,008	9,743	41·1
Glasgow . .	17,529	39,764	44·1

Can the difference observed in the percentage of girls of medium hair-colour have arisen solely through fluctuations of sampling?

In the two towns together the percentage of girls with medium hair-colour is 43·5 per cent. If this were the true percentage, the standard error of sampling for the difference between percentages observed in samples of the above sizes would be :

$$\epsilon_{12} = (43·5 \times 56·5)^{\frac{1}{2}} \times \left(\frac{1}{9743} + \frac{1}{39,764}\right)^{\frac{1}{2}}$$

$$= 0·56 \text{ per cent.}$$

The actual difference is 3·0 per cent., or over 5 times this, and could not have arisen through the chances of simple sampling.

If we assume that the difference is a real one and calculate the standard error by equation (19.5), we arrive at the same value, viz. 0·56 per cent. With such large samples the difference could not, accordingly, be obliterated by the fluctuations of simple sampling alone.

19.29. *Case 3.*—Two samples are drawn from distinct material or different universes, as in the last case, giving proportions of A's p_1 and p_2, but in lieu of comparing the proportion p_1 with p_2 it is compared with the proportion of A's in the two samples together, viz. p_0, where, as before,

$$p_0 = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Required to find whether the difference between p_1 and p_0 can have arisen as a fluctuation of simple sampling, p_0 being the true proportion of A's in both samples.

This case corresponds to the testing of an association which is indicated by a comparison of the proportion of A's amongst the B's with the proportion of A's in the universe. The general treatment is similar to that of Case 2, but the work is complicated owing to the fact that errors in p_1 and p_0 are not independent.

If ϵ_{01} be the standard error of the difference between p_1 and p_0, we have at once :

$$\epsilon_{01}^2 = \epsilon_0{}^2 + \epsilon_1{}^2 - 2r_{01}\epsilon_0\epsilon_1$$

$$= p_0 q_0 \left\{\frac{1}{n_1 + n_2} + \frac{1}{n_1} - 2r_{01}\frac{1}{\sqrt{n_1}}\frac{1}{\sqrt{n_1 + n_2}}\right\}$$

r_{01} being the correlation between errors of simple sampling in p_1 and p_0. But from the above equation relating p_0 to p_1 and p_2, writing it in terms

of deviations in p_0, p_1 and p_2, multiplying by the deviation in p_1 and summing, we have, since errors in p_1 and p_2 are uncorrelated :

$$r_{01} = \frac{n_1}{n_1 + n_2} \cdot \frac{\epsilon_1}{\epsilon_0} = \sqrt{\frac{n_1}{n_1 + n_2}}$$

Therefore finally :

$$\epsilon_{01}^2 = \frac{p_0 q_0}{n_1 + n_2} \cdot \frac{n_2}{n_1} \qquad . \qquad . \qquad . \qquad . \quad (19.6)$$

Unless the difference between p_0 and p_1 exceed, say, some three times this value of ϵ_{01}, it may have arisen solely by the chances of simple sampling..

It will be observed that if n_1 be very small compared with n_2, ϵ_{01} approaches, as it should, the standard error for a sample of n_1 observations.

We omit, in this case, the allied problem whether, if the difference between p_1 and p_0 indicated by the samples were real, it might be wiped out in other samples of the same size by fluctuations of simple sampling alone. The solution is a little complex, as we no longer have $\epsilon_0^2 = p_0 q_0 / (n_1 + n_2)$.

Example 19.8.—Taking now the figures of Example 19.7, suppose that we had compared the proportion of girls of medium hair-colour in Edinburgh with the proportion in Glasgow and Edinburgh together. The former is 41·1 per cent., the latter 43·5 per cent., difference 2·4 per cent. The standard error of the difference between the percentages observed in the sub-sample of 9743 observations and the entire sample of 49,507 observations is, therefore,

$$\epsilon_{01} = (43 \cdot 5 \times 56 \cdot 5)^{\frac{1}{2}} \left(\frac{39,764}{49,507 \times 9743} \right)^{\frac{1}{2}} = 0 \cdot 45 \text{ per cent.}$$

The actual difference is over five times this (the ratio must, of course, be the same as in Example 19.7), and could not have occurred as a mere error of sampling.

Effect of Removing the Limitations of Simple Sampling.

19.30. Let us now consider the effect on the standard error of the removal of the conditions of simple sampling which we discussed in **19.19 to 19.24.**

The breakdown of the condition we discussed in **19.20**, namely, that the proportion of A's in the universe should remain constant for all samples, might occur if we took a number of samples from a changing universe or from different strata of a universe which was not homogeneous.

We may represent such circumstances in a case of artificial chance by supposing that for the first f_1 throws of n dice the chance of success for each die is p_1, for the next f_2 throws p_2, for the next f_3 throws p_3, and so on, the chance of success varying from time to time, just as the chance of death, even for individuals of the same age and sex, varies from district to district. Suppose, now, that the records of all these throws are pooled together. The mean number of successes per throw of the n dice is given by

$$M = \frac{n}{N}(f_1 p_1 + f_2 p_2 + f_3 p_3 + \; . \; . \; .) = n p_0$$

where $N = S(f)$ is the whole number of throws, and p_0 is the mean value $S(fp)/N$ of the varying chance p. To find the standard deviation of the number of successes at each throw, consider that the first set of throws contributes to the sum of the squares of deviations an amount

$$f_1[np_1q_1 + n^2(p_1 - p_0)^2]$$

np_1q_1 being the square of the standard deviation for these throws, and $n(p_1 - p_0)$ the difference between the mean number of successes for the first set and the mean for all the sets together. Hence the standard deviation σ of the whole distribution is given by the sum of all quantities like the above, or

$$N\sigma^2 = nS(fpq) + n^2S\{f(p - p_0)^2\}$$

Let σ_p be the standard deviation of p, then the last sum is $Nn^2\sigma_p^2$, and substituting $1 - p$ for q, we have:

$$\sigma^2 = np_0 - np_0^2 - n\sigma_p^2 + n^2\sigma_p^2$$
$$= np_0q_0 + n(n-1)\sigma_p^2 \quad . \qquad . \qquad . \qquad . \quad (19 \cdot 7)$$

This is the formula corresponding to equation (19.1); if we deal with the standard deviation of the *proportion* of successes, instead of that of the absolute number, we have, dividing through by n^2, the formula corresponding to equation (19.2), viz.

$$s^2 = \frac{p_0q_0}{n} + \frac{n-1}{n}\sigma_p^2 \quad . \qquad . \qquad . \qquad . \quad (19.8)$$

19.31. If n be large and s_0 be the standard error calculated from the mean proportion of successes p_0, equation (19.8) is sensibly of the form

$$s^2 = s_0^2 + \sigma_p^2$$

We have thus analysed s^2 into two parts, s_0^2 the portion due to deviations from the mean p_0, and σ_p^2 the portion due to variations of the p's about their mean. The former we may regard as the contribution to s^2 due to chance fluctuations; the latter as the contribution due to real variation of the proportions among the different strata of the universe.

In conformity with later work we shall continue to call s (or σ if we are dealing with frequencies) the standard error, although the sampling is no longer simple. The deviation s is still, in fact, the standard deviation of the various sample values of p about the mean value. The term s_0 (or $\sqrt{np_0q_0}$), on the other hand, is what the standard error would have been if the sampling had been simple, and from the above equation we accordingly see that the effect of the breakdown of the first condition for simple sampling is to increase the standard error.

The values of $\sqrt{s^2 - s_0^2}$ are tabulated at the foot of Table 19.1, which shows data relating to the deaths of women in childbirth in certain groups of districts.

The values of $\sqrt{s^2 - s_0^2}$ suggest an almost uniform value of σ_p, about 0·8, in the deaths of women per 1000 births, *i.e.* that in each of the categories " number of births in the decade " there is real variability in the chances of individual women succumbing.

TABLE 19.1.—*Showing Frequencies of Registration Districts in England and Wales with Different Proportions of Deaths in Childbirth (including Deaths from Puerperal Fever) per 1000 Births in the Same Year.* (Data from Decennial Supplement to Fifty-fifth Annual Report of Registrar-General for England and Wales. Decade 1881–90.)

Deaths in Childbirth per 1000 Births.	Number of Births in the Decade.						
	1500 to 2500.	3500 to 4000.	4500 to 5000.	10,000 to 15,000.	15,000 to 20,000.	30,000 to 50,000.	50,000 to 90,000.
1·5– 2·0	—	—	2	—	—	—	—
2·0– 2·5	1	—	1	1	—	—	—
2·5– 3·0	1	3	1	—	—	—	—
3·0– 3·5	1	5	2	4	—	1	2
3·5– 4·0	5	6	5	8	5	5	9
4·0– 4·5	6	5	8	23	4	9	6
4·5– 5·0	2	5	9	14	11	7	5
5·0– 5·5	7	3	6	14	6	8	7
5·5– 6·0	5	3	4	5	2	5	4
6·0– 6·5	1	5	1	—	4	1	1
6·5– 7·0	3	1	1	3	—	2	1
7·0– 7·5	1	1	—	—	—	4	—
7·5– 8·0	—	—	—	—	—	1	—
8·0– 8·5	—	—	—	—	—	—	—
8·5– 9·0	1	1	—	—	1	—	—
9·0– 9·5	—	—	—	—	—	—	—
9·5–10·0	1	—	—	1	—	—	—
10·0–10·5	—	—	—	—	—	—	—
10·5–11·0	1	—	—	—	—	—	—
Total	36	38	40	73	33	43	35
Mean	5·29	4·71	4·45	4·68	4·99	5·13	4·64
Standard deviation	1·77	1·37	1·09	1·01	0·99	1·12	0·87
Theoretical standard deviation corresponding to mean births	1·62	1·12	0·97	0·61	0·53	0·36	0·26
$\sqrt{s^2 - s_0^2}$	0·71	0·80	0·51	0·80	0·84	1·07	0·83

The figures of this case also bring out clearly one important consequence of (19.8), viz. that if we make n large, s becomes sensibly equal to σ_p, while if we make n small, s becomes more nearly equal to $p_0 q_0/n$. Hence, if we want to know the significant standard deviation of the proportion p —the measure of its fluctuation owing to definite causes—n should be made as large as possible ; if, on the other hand, we want to obtain good illustrations of the theory of simple sampling, n should be made small. If n be very large, the actual standard error may evidently become almost indefinitely large compared with the standard deviation of *simple* sampling. Thus during the twenty years 1855–74 the death-rate in England and Wales fluctuated round a mean value of 22·2 per thousand with a standard

deviation (s) of 0·86. Taking the mean population as roughly 21 millions, the standard deviation of simple sampling (s_0) is approximately

$$\sqrt{\frac{22 \times 978}{21 \times 10^6}} = 0\cdot032 \text{ per thousand}$$

This is only about one twenty-seventh of the actual value.

19.32. Now consider the effect of altering the second condition of simple sampling dealt with in **19.21**, viz. the circumstances that regulate the appearance of the character observed shall be the same for every individual or every sub-class in each of the universes from which samples are drawn. Suppose that in a group of n dice thrown the chances for m_1 dice are $p_1 q_1$; for m_2 dice, $p_2 q_2$, and so on, the chances varying for different dice, but being constant throughout the experiment. The case differs from the last, as in that the chances were the same for every die, at any one throw, but varied from one throw to another; now they are constant from throw to throw, but differ from one die to another as they would in any ordinary set of badly made dice. Required to find the effect of these differing chances.

For the mean number of successes we evidently have:

$$M = m_1 p_1 + m_2 p_2 + m_3 p_3 + \ldots$$

$$= n p_0$$

p_0 being the mean chance $S(mp)/n$. To find the standard deviation of the number of successes at each throw, it should be noted that this may be regarded as made up of the number of successes in the m_1 dice for which the chances are $p_1 q_1$, together with the number of successes amongst the m_2 dice for which the chances are $p_2 q_2$, and so on; and these numbers of successes are all independent. Hence,

$$\sigma^2 = m_1 p_1 q_1 + m_2 p_2 q_2 + m_3 p_3 q_3 + \ldots$$

$$= S(mpq)$$

Substituting $1 - p$ for q, as before, and using σ_p to denote the standard deviation of p,

$$\sigma^2 = n p_0 q_0 - n \sigma_p^2 \qquad \cdot \qquad \cdot \qquad \cdot \qquad \cdot \quad (19.9)$$

or if s be, as before, the standard error of the *proportion* of successes,

$$s^2 = \frac{p_0 q_0}{n} - \frac{\sigma_p^2}{n} \qquad \cdot \qquad \cdot \qquad \cdot \quad (19.10)$$

Hence, in this case the standard error s is less than the standard error of simple sampling.

19.33. The extent to which the standard error is affected may conceivably be considerable. To take a limiting case, if p be zero for half the events and unity for the remainder, $p_0 = q_0 = \frac{1}{2}$, and $\sigma_p = \frac{1}{2}$, so that s is zero. To take another illustration, still somewhat extreme, if the values of p are uniformly distributed over the whole range between 0 and 1, $p_0 = q_0 = \frac{1}{2}$ as before, but $\sigma_p^2 = 1/12 = 0\cdot0833$ (**8.14**, p. 143). Hence, $s^2 = 0\cdot1667/n$, $s = 0\cdot408/\sqrt{n}$, instead of $0\cdot5/\sqrt{n}$, the value of s if the chances are $\frac{1}{2}$ in every

case. In most practical cases, however, the effect will be much less.   Thus the standard deviation of simple sampling for a death-rate of, say, 12 per thousand in a population of uniform age and one sex is $(12 \times 988)^{\frac{1}{2}}/\sqrt{n}$ $=109/\sqrt{n}$.   In a population of the age composition of that of England and Wales, however, the death-rate is not, of course, uniform, but varies from a high value in infancy (say 64 per thousand), through very low values (2 to 3 per thousand) in childhood to continuously increasing values in old age ; the standard deviation of the rate within such a population is roughly about 24 per thousand.   But the effect of this variation on the standard deviation of simple sampling is quite small, for, as calculated from equation (19.10),

$$s^2 = \frac{1}{n}(12 \times 988 - 576)$$

$$s = 106/\sqrt{n}$$

as compared with $109/\sqrt{n}$.

19.34. We have, finally, to pass to the condition referred to in **19.23**, and to discuss the effect of a certain amount of dependence between the several " events " in each sample.   We shall suppose, however, that the two other conditions are fulfilled, the chances p and q being the same for every event at every trial, and constant throughout the experiment.   The standard deviation for each event is $(pq)^{\frac{1}{2}}$ as before, but the events are no longer independent; instead, therefore, of the simple expression

$$\sigma^2 = npq$$

we must have (*cf.* **16.2**, p. 297)

$$\sigma^2 = npq + 2pq(r_{12} + r_{13} + \ldots + r_{23} + \ldots)$$

where r_{12}, r_{13}, etc. are the correlations between the results of the first and second, first and third events, and so on—correlations for variables (number of successes) which can only take the values 0 and 1, but may nevertheless be treated as ordinary variables.   There are $n(n-1)/2$ correlation coefficients, and if, therefore, r is the arithmetic mean of the correlations, we may write :

$$\sigma^2 = npq[1 + r(n-1)] \qquad . \qquad . \qquad . \quad (19.11)$$

The standard deviation of simple sampling will therefore be increased or diminished according as the average correlation between the results of the single events is positive or negative, and the effect may be considerable, as σ may be reduced to zero or increased to $n(pq)^{\frac{1}{2}}$.   For the standard deviation of the proportion of successes in each sample we have the equation

$$s^2 = \frac{pq}{n}[1 + r(n-1)] \qquad . \qquad . \qquad . \quad (19.12)$$

19.35. It should be noted that, as the means and standard deviations . for our variables are all identical, r is the correlation coefficient for a table formed by taking all possible pairs of results in the n events of each sample.

It should also be noted that the case when r is positive covers the departure from the rules of simple sampling discussed in **19.30–19.31** ;

for if we draw successive samples from different records, this introduces the positive correlation at once, even although the results of the events *at each trial* are quite independent of one another. Similarly, the case discussed in **19.32–19.33** is covered by the case when r is negative; for if the chances are not the same for every event at each trial, and the chance of success for some one event is above the average, the mean chance of success for the remainder must be below it. The present case is, however, best kept distinct from the other two, since a positive or negative correlation may arise for reasons quite different from those discussed in **19.30–19.33**.

19.36. As a simple illustration, consider the important case of sampling from a limited universe, *e.g.* of drawing n balls in succession from the whole number w in a bag containing pw white balls and qw black balls. On repeating such drawings a large number of times, we are evidently equally likely to get a white ball or a black ball for the first, second or nth ball of the sample; the correlation table formed from all possible pairs of every sample will therefore tend in the long run to give just the same form of distribution as the correlation table formed from all possible pairs of the w balls in the bag. But from **13.32**, page 257, we know that the correlation coefficient for this table is $-1/(w-1)$, whence

$$\sigma^2 = npq\left(1 - \frac{n-1}{w-1}\right)$$

$$= npq\frac{w-n}{w-1}$$

If $n = 1$, we have the obviously correct result that $\sigma = (pq)^{\frac{1}{2}}$, as in drawing from unlimited material; if, on the other hand, $n = w$, σ becomes zero as it should, and the formula is thus checked for simple cases. For drawing 2 balls out of 4, σ becomes $0{\cdot}816(npq)^{\frac{1}{2}}$; for drawing 5 balls out of 10, $0{\cdot}745(npq)^{\frac{1}{2}}$; in the case of drawing half the balls out of a very large number, it approximates to $(0{\cdot}5npq)^{\frac{1}{2}}$, or $0{\cdot}707(npq)^{\frac{1}{2}}$.

19.37. In the case of contagious or infectious diseases, or of certain forms of accident that are apt, if fatal at all, to result in wholesale deaths, r is positive, and if n be large (as it usually is in such cases), a very small value of r may easily lead to a very great increase in the observed standard deviation. It is difficult to give a really good example from actual statistics, as the conditions are hardly ever constant from one year to another, but the following will serve to illustrate the point. During the twenty years 1887–1906 there were 2107 deaths from explosions of firedamp or coal-dust in the coal-mines of the United Kingdom, or an average of 105 deaths per annum. From **19.15** it follows that this should be the square of the standard deviation of simple sampling, or the standard deviation itself approximately 10·3. But the square of the actual standard deviation (the standard error) is 7178, or its value 84·7, the numbers of deaths ranging between 14 (in 1903) and 317 (in 1894). This large standard deviation, to judge from the figures, is partly, though not wholly, due to a general tendency to decrease in the numbers of deaths from explosions in spite of a large increase in the number of persons employed; but even if we ignore this, the magnitude of the standard deviation can be accounted for by a very small value of the correlation r, expressive of the fact that if an explosion is sufficiently serious to be fatal to one individual, it will probably

be fatal to others also. For if σ_0 denote the standard deviation of simple sampling, σ the standard deviation of sampling given by equation (19.11), we have :

$$r = \frac{\sigma^2 - \sigma_0^2}{(n-1)\sigma_0^2}$$

Whence, from the above data, taking the numbers of persons employed underground at a rough average of 560,000,

$$r = \frac{7073}{560,000 \times 105} = +0.00012$$

19.38. Summarising the preceding paragraphs, **19.30–19.37,** we see that if the chances p and q differ for the various universes, districts, years, materials, or whatever they may be from which the samples are drawn, the standard deviation observed (the standard error) will be greater than the standard deviation of simple sampling, as calculated from the average values of the chances ; if the average chances are the same for each universe from which a sample is drawn, but vary from individual to individual or from one sub-class to another within the universe, the standard deviation observed (the standard error) will be less than the standard deviation of simple sampling as calculated from the mean values of the chances; finally, if p and q are constant, but the events are no longer independent, the observed standard deviation (the standard error) will be greater or less than the simplest theoretical value according as the correlation between the results of the single events is positive or negative.   These conclusions further emphasise the need for caution in the use of standard errors.   If we find that the standard deviation in some case of sampling exceeds the standard deviation of simple sampling, two interpretations are possible : *either* that p and q are different in the various universes from which samples have been drawn (*i.e.* that the variations are more or less significant), *or* that the results of the events are positively correlated *inter se.*   If the actual standard deviation fall short of the standard deviation of simple sampling two interpretations are again possible : *either* that the chances p and q vary for different individuals or sub-classes in each universe, while approximately constant from one universe to another, *or* that the results of the events are negatively correlated *inter se.*   Even if the actual standard deviation approaches closely to the standard deviation of simple sampling, it is only a conjectural and not a necessary inference that all the conditions of " simple sampling " are fulfilled.   Possibly, for example, there may be a positive correlation r between the results of the different events, masked by a variation of the chances p and q in sub-classes of each universe.

An Alternative Approach.

19.39.   The results of this chapter have been studied from a rather different point of view by a continental school of statisticians, among whose names those of Lexis and Charlier are prominent.

Lexis considers a number of samples of n individuals in which the proportions of successes observed are $p_1, p_2, \ldots p_s$, and sets himself to investigate the nature of the universe from which they were drawn— whether it is homogeneous and the samples may be regarded as obtained by simple sampling, whether it varies in time or place so that the samples

are not simple, and so on. He takes p to be the mean of the observed values $p_1 \ldots p_N$, and writes :

$$r = 0{\cdot}67449 \sqrt{\frac{pq}{n}}$$

He then defines

$$R = 0{\cdot}67449 \sqrt{\frac{S(p_k - p)^2}{N - 1}}$$

where the summation extends over all values of $p_1 \ldots p_N$, and writes

$$Q = \frac{R}{r}$$

19.40. Now, if the sampling is simple we may, in large samples, take the mean p to be an estimate of the true value, and r to be an estimate of the probable error of simple sampling of p. Also, we may take the quantity R to be an estimate of the probable error of p (see **23.5**).

Hence, for large samples, R is approximately equal to r, and $Q = 1$. This case, which is what we have called simple sampling, Lexis calls " normal dispersion."

19.41. On the other hand, if the universe is not constant while the samples are drawn, or if they come from different parts of a patchy universe, we get the case discussed in **19.30**. R is no longer an estimate of the probable error of a constant p, but may be split into two parts, one due to the sampling fluctuations of the observed values of p round the mean value, the other due to the variations of the true values round that mean. R will therefore be greater than r, as may be seen from equation (19.8), and $Q > 1$. This case Lexis calls " supernormal dispersion."

19.42. Similarly, in the case discussed in **19.32** we get R less than r, and hence $Q < 1$. This case Lexis calls " subnormal dispersion," and speaks of the data which give rise to it as " constrained " (*gebundene*).

The quantity Q is analogous to a quantity χ^2, which we shall consider at some length in Chapter 22 in discussing the significance of the deviations of observed frequencies from theoretical expectation.

SUMMARY.

1. Under simple sampling conditions, the proportion of successes in a sample may be taken as an estimate of the proportion of successes in the parent universe.

2. If p is the proportion of successes in the universe, the standard error of simple sampling of the *number* of successes is given by

$$\sigma = \sqrt{npq}$$

and of the *proportion* of successes by

$$s = \sqrt{\frac{pq}{n}}$$

3. The probability that an observed number of successes deviates from the expected number by more than three times the standard error is very

small. This fact enables us to set limits to the range within which the observed frequency lies when we know the theoretical frequency.

4. For large samples, the observed frequency of successes may be used to calculate the standard error, and this fact enables us to set limits to the range within which the theoretical frequency lies when we know the observed frequency.

5. For several samples, if the chance of success varies from sample to sample but remains constant within a sample, the standard error of the *number* of successes is given by

$$\sigma^2 = np_0q_0 + n(n-1)\sigma_p{}^2$$

and of the *proportion* of successes by

$$s^2 = \frac{p_0q_0}{n} + \frac{n-1}{n}\sigma_p{}^2$$

where p_0 is the mean of the varying chance of success, σ_p is the standard deviation of p, and n is the number of individuals in each sample.

If n is large and s_0 is the standard deviation calculated from the mean p_0, this last equation is approximately

$$s^2 = s_0{}^2 + \sigma_p{}^2$$

6. If the chance of success varies between the individuals of a sample but does not vary as between the different samples,

$$\sigma^2 = np_0q_0 - n\sigma_p{}^2$$

$$s^2 = \frac{p_0q_0}{n} - \frac{\sigma_p{}^2}{n}$$

7. If the chance of success remains constant for each member of each sample, but the events are not independent,

$$\sigma^2 = npq\{1 + r(n-1)\}$$

$$s^2 = \frac{pq}{n}\{1 + r(n-1)\}$$

where r is the mean of the correlations between the results of the events.

EXERCISES.

19.1. (Ref. (398): total of columns of all the 13 tables given.)
Compare the actual with the theoretical mean and standard deviation for the following record of 6500 throws of 12 dice, 4, 5 or 6 being reckoned as a "success" :—

Successes.	Frequency.	Successes.	Frequency.
0	·1	7	1351
1	14	8	844
2	103	9	391
3	302	10	117
4	711	11	21
5	1231	12	3
6	1411		
		Total	6500

19.2. (Quetelet, "Lettres . . . sur la théorie des probabilités.")

Balls were drawn from a bag containing equal numbers of black and white balls, each ball being returned before drawing another. The records were then grouped by counting the number of black balls in consecutive 2's, 3's, 4's, 5's, etc. The following are the distributions so derived for grouping by 5's, 6's, and 7's. Compare actual with theoretical means and standard deviations.

Successes.	(a) Grouping by Fives.	(b) Grouping by Sixes.	(c) Grouping by Sevens.
0	30	17	9
1	125	65	34
2	277	166	104
3	224	192	151
4	136	166	148
5	27	69	95
6	—	8	40
7	—	—	4
Total	819	683	585

19.3. The proportion of successes in the data of Exercise 19.1 is 0·5097. Find the standard deviation of the proportion with the given number of throws, and state whether you would regard the excess of successes as probably significant of bias in the dice.

19.4. In the 4096 drawings on which Exercise 19.2 is based 2030 balls were black and 2066 white. Is this divergence probably significant of bias?

19.5. (Data from Report I, Evolution Committee of the Royal Society, p. 17.) In breeding certain stocks, 408 hairy and 126 glabrous plants were obtained. If the expectation is one-fourth glabrous, is the divergence significant, or might it have occurred as a fluctuation of sampling?

19.6. 400 eggs are taken at random from a large consignment, and 50 are found to be bad. Estimate the percentage of bad eggs in the consignment and assign limits within which the percentage probably lies.

19.7. In a certain association table (data from Exercise 3.5) the following frequencies were obtained : —

$$(AB)=309, \quad (A\beta)=214, \quad (aB)=132, \quad (a\beta)=119$$

Can the association of the table have arisen as a fluctuation of simple sampling, the true association being zero?

19.8. The sex ratio at birth is sometimes given by the ratio of male to female births, instead of the proportion of male to total births. If Z is the ratio, $i.e.$ $Z=p/q$, show that the standard error of Z is approximately $(1+Z)\sqrt{\dfrac{Z}{n}}$, n being large, so that deviations are small compared with the mean.

19.9. In a random sample of 500 persons from town A, 200 are found to be consumers of cheese. In a sample of 400 from town B, 200 are also found to be consumers of cheese. Discuss the question whether the data reveal a significant difference between A and B so far as the proportion of cheese-consumers is concerned.

19.10. In a newspaper article of 1600 words in English 36 per cent. of the words are found to be of Anglo-Saxon origin. Assuming that simple sampling conditions hold, estimate the proportion of Anglo-Saxon words in the writer's vocabulary and assign limits to that proportion.

Suggest possible causes which might break down the three conditions for simple sampling.

19.11. If a series of random samples of different sizes is taken from the same material, show that the standard deviation of the observed proportions of successes in such sets is s, where

$$s^2 = \frac{pq}{H}$$

and H is the harmonic mean of the numbers in the samples.

19.12. Apply the result of the previous exercise to the following data (A. D. Darbishire, *Biometrika*, vol. 3, p. 30), giving percentages to the nearest unit of albinos obtained in 121 litters from hybrids of Japanese waltzing mice by albinos, crossed *inter se* :—

Percentage.	Frequency.	Percentage.	Frequency.
0	40	40	3
14	4	43	2
17	9	50	16
20	9	57	1
22	1	60	3
25	10	67	4
29	3	80	1
33	13	100	2

Calculate the actual standard deviation and compare it with the result given by the formula of the previous exercise. The expected proportion of albinos is 25 per cent., and the sizes of the litters are given in Example 7.5, page 130.

19.13. In a case of mice-breeding (see reference above) the harmonic mean number in a litter was 4·735, and the expected proportion of albinos 50 per cent. Find the standard deviation of simple sampling for the proportion of albinos in a litter, and state whether the actual standard deviation (21·63 per cent.) probably indicates any *real* variation, or not.

19.14. In the data of Table 11.6, page 202, the standard deviation of the proportion of male births per 1000 of all births is 7·46 and the mean proportion of male births 509·2. The harmonic mean number of births in a district is 5070. Find the significant standard deviation σ_p.

19.15. If for one half of n events the chance of success is p and the chance of failure q, whilst for the other half the chance of success is q and the chance of failure p, what is the standard deviation of the number of successes, the events being all independent?

19.16. The following are the deaths from smallpox during the twenty years 1882–1901 in England and Wales: —

1882	1317	1892	431
83	957	93	1457
84	2234	94	820
85	2827	95	223
86	275	96	541
87	506	97	25
88	1026	98	253
89	23	99	174
90	16	1900	85
91	49	1901	356

The death-rate from smallpox being very small, the rule of **19.15** may be applied to estimate the standard deviation of simple sampling. Assuming that the excess of the actual standard deviation over this can be entirely accounted for by a correlation between the results of exposure to risk of the individuals composing the population, estimate r. The mean population during the period may be taken in round numbers as 29 millions.

THE SAMPLING OF VARIABLES—LARGE SAMPLES.

Sampling of Variables.

20.1. We are now able to proceed from the sampling of attributes to the sampling of variables. Whereas in the last chapter we were interested in the question whether a member of a sample did or did not exhibit a particular attribute, we now have to study individuals which may take any of the values of a variable. It will no longer be possible, therefore, for us to classify each member of a sample under one of two heads, success or failure ; in general the values of the variate given by different trials will be spread over a range, which may be unlimited, limited by practical considerations, as in the case of height in human beings, or limited by theoretical considerations, as in the case of the correlation coefficient, which cannot lie outside the range $+1$ to -1.

20.2. To give concreteness to our discussions we shall occasionally find it useful to consider the sampling of variables as a kind of ticket sampling. We may picture our universe as made up of tickets, each bearing a recorded value of some variable X. Sampling may then be imagined to consist of the drawing of tickets and the noting of the values of X which they bear. In the great majority of cases with which we shall deal, X may have any value over a continuous range, and the ticket universe is to be conceived as being actually or practically infinite.

20.3. As in the case of attributes, our principal objects in studying these samples will be (a) to compare observation with expectation and to see how far deviations of one from the other can be attributed to fluctuations of sampling ; (b) to estimate from samples some characteristic of the parent, such as the mean of a variate ; and (c) to gauge the reliability of our estimates.

In order to grasp satisfactorily the ideas and assumptions upon which work of this kind is based, it is necessary to develop some theoretical considerations which have already been touched upon in the last chapter. This we now proceed to do.

Sampling Distributions.

20.4. If we take a number of samples from a universe and calculate some function,[1] such as the mean or the standard deviation, of each sample, we shall in general get a series of different values, one for each sample. If the number of samples is at all large, these values may be grouped in a frequency distribution ; and as the number of samples becomes larger, this distribution will approach the " ideal " form of a continuous curve. Such a distribution is called a **sampling distribution.**

[1] Quantities such as means, standard deviations, moments, correlation coefficients and so forth will be referred to generically as "parameters."

20.5. As an illustration, consider the universe of 8585 men, classified according to height, of Table 6.7, page 94. In Chapter 18 we showed how to draw a random sample of 10 individuals from this universe, and for one sample we calculated the mean. The following table shows the 100 values of the sample mean obtained by taking 100 such samples arranged in the form of a frequency table : —

TABLE 20.1.—*Frequency Distribution of Means of Samples of 10 from the Universe of the last column of Table 6.7, page 94.*

Value of Mean in Sample (inches) less $\frac{1}{16}$ inch.	Number of Samples with Specified Values of the Mean.
64·4–	1
64·8–	—
65·2–	1
65·6–	11
66·0–	12
66·4–	16
66·8–	22
67·2–	18
67·6–	14
68·0–	4
68·4–	1
Total	100

This distribution is not very regular, owing to the smallness of the total frequency.

20.6. As a second illustration we take some data obtained by random sampling with Tippett's numbers from a bivariate normal universe with correlation +0·9. 500 samples of 10 were taken and the correlation coefficient of each sample worked out. The frequency distribution of the 500 values was as follows (data adapted from P. R. Rider, " Distribution of Correlation Coefficient in Small Samples," *Biometrika*, vol. 24, 1932, p. 382) :—

TABLE 20.2.—*Frequency Distribution of Correlation Coefficients in Samples of 10 from a Normal Universe.*

Value of r in Sample.	Frequency.
– 0·1–0·0	2
0·0–0·1	0
0·1–0·2	0
0·2–0·3	2
0·3–0·4	4
0·4–0·5	7
0·5–0·6	30
0·6–0·7	44
0·7–0·8	102
0·8–0·9	178
0·9–1·0	131
Total	500

Here the distribution is more regular, the number of samples being five times as large. In general we expect that as the number of samples increases, the distribution will tend more and more to a continuous curve.

Use of the Sampling Distribution.

20.7. Let us suppose that we are given the sampling distribution of a parameter, and that the frequency (y) may be represented in terms of the variate (x) by a continuous curve,

$$y = F(x)$$

The frequency with which a given value x_0 of the parameter occurs in a large number of samples will be represented by the ordinate of the curve at the point whose abscissa is x_0. We have had an example of this in the normal curve.

The number of samples which give a value of x greater than x_0 will be represented by the area to the right of the ordinate at x_0; the number giving a value less than x_0 will be represented by the remaining area to the left.

Hence, the chance that any sample chosen at random from all possible samples will give a value of x greater than x_0 is given by the area to the right of the ordinate at x_0 divided by the total area of the curve, which represents the total number of samples; and the chance that the sample will give a value of x less than x_0 is given by the area to the left of the ordinate of x_0 divided by the total area.

Similarly, the chance that a sample would give a value of x lying between, say, x_1 and x_2 is the area lying between the ordinates at the points x_1 and x_2 divided by the total area.

20.8. In **10.21** we referred to the fact that areas could be expressed in the notation of the integral calculus. In fact, we may write the area of the curve between x_1 and x_2 as

$$\int_{x_1}^{x_2} F(x)dx$$

and hence we may express P, the probability that a sample will give a value between x_1 and x_2, as

$$P = \int_{x_1}^{x_2} F(x)dx \bigg/ \int_{-\infty}^{\infty} F(x)dx$$

where we assume the extreme limits to be $\pm\infty$ as in the normal curve. In particular, the probability that the sample will give a value of x greater than x_0 is given by

$$P = \frac{\int_{x_0}^{\infty} F(x)dx}{\int_{-\infty}^{\infty} F(x)dx}$$

As a rule, we can choose our units so that the area of the curve is unity. This simplifies the above expressions; for the denominator, being equal to unity, may be omitted.

20.9. Now let us suppose that, knowing the form of the sampling distribution and hence being able to calculate P for any given x_0, we take a sample and find that it gives a very low value of P. We are then faced with three possibilities: either a very improbable event has occurred; or the assumptions on which we obtained the sampling distribution were incorrect; or there is something wrong with our sampling technique. Which of these explanations we adopt is to some extent a matter of choice, but if we have tested our sampling, or on other grounds have no reason to suspect it, we shall, as a rule, be led to query the hypotheses on which the sampling distribution was obtained.

This, in effect, is what we did in the previous chapter. It so happens that in the simple sampling of attributes we know that the exact form of the sampling distribution is $N(q + p)^n$, where p is the chance of success. Without examining this distribution too closely we can say that only a very small part of it lies outside the range $\pm 3\sigma$. Hence, if we find a sample giving a value outside the range $\pm 3\sqrt{npq}$, we suspect the hypothesis on which the distribution was based; and this, unless we prefer to suppose that our sampling was not in fact simple, leads us to suspect the value of p, which completely determines the sampling distribution.

20.10. In the previous chapter we regarded the probability of a sample giving a value differing by more than 3σ from the mean value as so remote that in every case we should be justified in looking for some definite cause of the discrepancy. This is only a conventional range, based upon the empirical fact that in most single-humped universes it includes nearly all the members; but it is a convenient one to take and we shall use it again below. For certain purposes, however, we might be prepared to use a narrower range which, though not giving such a small probability that a sample lay outside it, yet indicated considerable improbability in the divergence of observation from expectation, and enabled us to criticise the validity of our hypotheses with some degree of assurance. We give one or two examples below.

20.11. In practice nearly all the sampling distributions we have to consider are based on simple sampling. It is therefore convenient to speak briefly of a " sampling distribution," meaning thereby a sampling distribution obtained under simple (and random) conditions.

Example 20.1.—The sampling distribution of a parameter is a normal universe with mean 3 units and standard deviation 2 units. What is the probability that a sample will give a value of the parameter greater than 6 units ?

Here the value 6 is three units, *i.e.* $1 \cdot 5\sigma$, to the right of the mean. The required probability is therefore the area of the normal curve to the right of an ordinate $1 \cdot 5\sigma$ to the right of the mean, divided by the total area of the curve.

This ratio can be obtained at once from Table 2 of the Appendix. We see, in fact, that the greater fraction of the area of the curve corresponding to $\frac{x}{\sigma} = 1 \cdot 5$ is $0 \cdot 93319$. The smaller fraction is therefore $0 \cdot 06681$, which gives us the required probability.

Example 20.2.—If the sampling distribution of a parameter is normal, with zero mean and standard deviation σ, what is the value of the

parameter such that the chances are 99 to 1 against a sample giving a value in excess of that value ?

We have to find x such that the area of the curve to the right of the ordinate at x is 0·01, or the area to the left 0·99.

From Appendix Table 2 :

$$\text{If } \frac{x}{\sigma} = 2\cdot3, \text{ greater fraction of area} = 0\cdot98928$$

$$\text{and if } \frac{x}{\sigma} = 2\cdot4 \quad ,, \qquad ,, \qquad ,, \quad = 0\cdot99180$$

Hence, by simple interpolation the greater fraction is 0·99 if $\frac{x}{\sigma} = 2\cdot33$ approximately, and hence the required value is $2\cdot33\sigma$.

Example 20.3.—It very frequently happens in sampling inquiries that we are interested in the probability that a sample value exceeds a given value x_0 *in absolute value, i.e.* that it is greater than x_0 or less than $-x_0$. We can ascertain this probability without much trouble from the ordinary table of areas of the normal curve if the distribution is normal. Consider, for instance, the data of Example 20.1. Here we found the probability that a sample would give a value greater than $1\cdot5\sigma$. If we want the probability that it would give a value greater than $1\cdot5\sigma$ in absolute value, we have :

$$P = \text{Area to right of ordinate at } 1\cdot5\sigma$$
$$+ \text{Area to left of ordinate at } -1\cdot5\sigma$$

Since the curve is symmetrical, the two areas in question are equal, and

$$P = 2(1 - 0\cdot93319)$$
$$= 0\cdot13362$$

For convenience, however, we have given in Table 3 of the Appendix the values of this probability directly in terms of $\frac{x}{\sigma}$. From this table we have at once, for $\frac{x}{\sigma} = 1\cdot5$,

$$P = 0\cdot13361$$

the difference in the last place being due merely to our having multiplied . by 2 in the former value of P a quantity which was rounded up to the nearest figure, whereas P in the latter case was calculated more accurately.

20.12. To apply the results of **20.7** to **20.11** in practice for the purpose of discussing the universe from which the samples came, we require to know two things : (*a*) What is the relation between the sampling distribution and the parent distribution, and (*b*) what is the form, at least approximately, of the sampling distribution of a given parameter from a given universe ?

20.13. If the sampling is to be of much use in enabling us to estimate the value of a parameter in the parent, we should expect most of our estimates to be somewhere near the mark, and only comparatively few to be very far from the true value of the quantity estimated ; and further, we

expect that, in general, the further the estimates are from the truth the fewer there will be of them.

To put this more formally, we expect that the sampling distribution will have a peak somewhere close to the value of the parameter which corresponds to the true value in the parent. If it does not, the distribution is probably biased and our samples are likely to be misleading.

The first *desideratum* in our sampling is, therefore, that it shall not lead to a biased distribution. We have seen in Chapter 18 the difficulties of eliminating bias in the sampling process itself. Where, therefore, the more practical considerations alluded to in that chapter impose no limitation, we must use unbiased sampling ; and this means that our sampling must be random. In this connection it must be remembered that we cannot judge from the samples themselves whether the sampling is random or not, though we may suspect it. Separate tests, or the use of some accredited method, are to be recommended where practicable.

20.14. Knowledge of the form of the sampling distribution of a parameter, even of an approximate kind, is by no means easy to secure. We saw that in the case of the simple sampling of attributes it was possible to deduce the sampling distribution in an exact form. We are not always in this fortunate position here—in fact, rarely so. The principal difficulties are :

(*a*) The form of the parent universe frequently is unknown.

(*b*) Even if the form of the parent is known, certain of its constants may be unknown ; for instance, we may know that a universe is normal but be ignorant of its mean and standard deviation.

(*c*) If the parent is completely known, the form of the sampling distribution can be deduced theoretically in certain circumstances, and in particular if the sampling is simple ; but in practice the mathematical problems which arise usually are very complex, and even if they are tractable may be of no use owing to the enormous arithmetical labour involved in expressing a solution in serviceable form.

20.15. If the samples are small these difficulties are formidable, even for simple sampling. With large samples, however, we are able to make certain legitimate approximations and assumptions which greatly simplify the problem. For the rest of this chapter and in the next we shall be concerned solely with large samples.

Simple Sampling of Variables.

20.16. We shall also be thinking mainly in terms of simple sampling (**19.3**). It is unnecessary to recapitulate here the discussion of simple sampling which we gave in the previous chapter. The assumptions which we considered in **19.19** to **19.24** apply *mutatis mutandis* to the simple sampling of variables.

(*a*) We assume that we are drawing from precisely the same record during the whole of the sampling ; if we picture our parent universe as a card universe, the chance of drawing a card with any given value X is the same for each sample.

(*b*) We assume not only that we are drawing from the same record throughout, but that *each of our cards* at each drawing may be regarded quite strictly as drawn from the same record (or from identically similar

records) : *e.g.* if our card record is contained in a series of bundles, we must not make it a practice to take the first card from bundle number 1, the second card from bundle number 2, and so on, or else the chance of drawing a card with a given value of X, or a value within assigned limits, may not be the same for each individual card at each drawing.

(*c*) We assume that the drawing of each card is entirely independent of that of every other, so that the value of X recorded on card 1, at each drawing, is uncorrelated with the value of X recorded on card 2, 3, 4, and so on.   It is for this reason that we spoke of the record, in **20.2**, as containing a practically infinite number of cards, for otherwise the successive drawings at each sampling would not be independent : if the bag contains ten tickets only, bearing the numbers 1 to 10, and we draw the card bearing 1, the average of the following cards drawn will be higher than the mean of all cards drawn ; if, on the other hand, we draw the 10, the average of the following cards will be lower than the mean of all cards—*i.e.* there will be a negative correlation between the number on the card taken at any one drawing and the card taken at any other drawing.   Without making the number of cards in the bag indefinitely large, we can, as already pointed out · for the case of attributes, eliminate this correlation by replacing each card before drawing the next.

Approximations in the Theory of Large Samples.

20.17.   We can now consider the approximations which are possible in the theory of large samples.

In the first place, since we have supposed bias to be eliminated, the sample values of a parameter will be grouped about the true value, and if the samples are large, will differ by comparatively small quantities from that value.   Hence, we may take a sample value as an estimate of the true value.   That is to say, if we have a large sample (which may consist of a number of samples run together), we may calculate the parameter from it precisely as we should proceed if we were calculating the parameter for the universe as a whole, and take that value as our estimate. Thus, the mean of the sample may be taken as an estimate of the mean of the universe.

20.18.   This rule is not quite so obvious as it appears.   Suppose, for example, that we are estimating the standard deviation of a universe. In accordance with the previous paragraph we should take the standard deviation of the sample.   But in calculating this quantity we should have to use deviations, not from the true mean, but from the mean in the sample, which may differ from the true mean and to that extent affect the value of the estimate.   We shall, in fact, see later that if x_1, x_2, . . . x_n are the values in the sample and $\bar{x}$ their mean, there are reasons for preferring the estimate $s^2 = \dfrac{1}{n-1}\mathrm{S}(x - \bar{x})^2$ to the estimate $s^2 = \dfrac{1}{n}\mathrm{S}(x - \bar{x})^2$ for the variance.   If n is large, however, the difference is unimportant ; we can ignore it until we come to deal with small samples.

20.19.   Secondly, as in the case of attributes, we can use these estimates in calculating the constants of the sampling distribution, since they differ only by small quantities from the real values.   We saw, for instance, that we were justified in taking the value of p in a large sample

in calculating the standard deviation $\sqrt{npq}$ of the sampling distribution. We shall find that the standard deviation of the sampling distribution of the mean of samples from a normal universe involves the standard deviation of the parent; and in this case we can evaluate that quantity by using the standard deviation of the sample in place of the unknown standard deviation of the parent.

20.20. Finally, it is a very remarkable fact that the sampling distributions of many parameters, obtained under simple sampling conditions, tend for large samples to a single-humped form either exactly or very closely normal. The evidence for this statement is partly theoretical, partly experimental. It may be shown that, for simple samples from a normal universe, the sampling distributions of most parameters are exactly normal for large samples—some, in fact, are normal for small samples. Following up this work, a number of experiments has been carried out on universes which are not normal; and it appears that the parent can deviate quite markedly from the normal form without affecting the normality of the sampling distribution to any great extent provided, as before, that the samples are large.

In most of our work we shall not require to assume that the sampling distribution is normal. It will be sufficient to assume that a range of 3σ on each side of the mean includes the major portion of the distribution, and we can confidently take this to be so unless the parent exhibits very marked skewness.

20.21. It will now be apparent that the difficulties we specified in **20.14** have to a great extent been met. Provided that we know the parent distribution to be not unduly skew, we need not know its exact form; and the sampling distribution can be represented satisfactorily, if not exactly specified, by a mean and standard deviation which may be estimated from the data of the sample.

Standard Error.

20.22. As in the last chapter, we shall refer to the standard deviation of the sampling distribution as the standard error. In most cases we shall be dealing with simple sampling distributions, but it is convenient to use the term in this wider sense, although the word " error " is not altogether appropriate in some instances. In general, as we have seen, we are justified in taking a range of ± 3 times the standard error as determining limits outside which the value of the parameter given by a sample probably does not lie. We can therefore use the standard error, as we have already used it for attributes, to gauge the precision of an estimate or to permit a judgment being made of the divergence between expected and observed values.

In the remainder of this chapter, and in the next, we shall therefore be concerned mainly in finding expressions for the standard errors of the various parameters which we have to estimate. Their use we shall illustrate in examples as we go along. In certain cases we shall also consider the effect of a breakdown in the conditions of simple sampling.

Standard Error of a Percentile, Quartile and Median.

20.23. Let us first of all consider the case of percentiles, which is intimately related to that of attributes.

Consider the distribution of a variate X in an indefinitely large sample. (This is not necessarily the same as the distribution in the parent, owing to the possible presence of bias ; but if bias is excluded, and the sampling is simple, it is the same as the parent form.)

Let X_p be a value of X such that pN values of X in this distribution lie above it and qN below it. Thus, if the sampling is unbiased, $p = \frac{1}{10}$ would give us the upper decile in the indefinitely large sample, $p = \frac{1}{2}$ the median, and so on.

A sample of n will contain various values of X. Let the proportion of values above X_p be $p + \delta$; and let ϵ be the adjustment to be made in X_p so that the proportion of values of X above $X_p + \epsilon$ is p. The values δ and ϵ may be regarded as sampling fluctuations.

Considering now the sample of n, we have that

the proportion of values above X_p $= p + \delta$

,,   ,,   ,, $X_p + \epsilon = p$

Hence,

$$\delta = \text{proportion of values between } X_p \text{ and } X_p + \epsilon$$

Now if n be large, the proportion of values between X_p and $X_p + \epsilon$ in the sample will, to a close approximation, be the proportion of values between those quantities in the distribution of an indefinitely large sample. Consider then this distribution and let the standard deviation of X in it be σ. If we take the distribution as drawn to scale with unit standard deviation and unit area, the proportion of values between X_p and $X_p + \epsilon$ is the area of the curve between ordinates at the points $\frac{X_p}{\sigma}$ and $\frac{X_p + \epsilon}{\sigma}$.

Now if n be large, ϵ will be small, for the value of a parameter in the sample of n will lie close to the value in the indefinitely large sample.

Hence the area between $\frac{X_p}{\sigma}$ and $\frac{X_p + \epsilon}{\sigma}$ is approximately rectangular, and

if we call the $\frac{X_p}{\sigma}$ ordinate y_p, the area will be $y_p \times \frac{\epsilon}{\sigma}$.

Hence,

$$\delta = y_p \times \frac{\epsilon}{\sigma}$$

or

$$\epsilon = \frac{\sigma}{y_p} \delta$$

Now δ is the deviation of the observed proportions from the value p ; and from our study of attributes we know that the observed proportions $p + \delta$ will centre round the mean p with standard deviation $\sqrt{\frac{pq}{n}}$.

Hence δ centres round zero mean with standard deviation $\sqrt{\frac{pq}{n}}$. Since

ϵ bears a constant ratio $\dfrac{\sigma}{y_p}$ to δ, it follows that ϵ will be distributed about zero mean with standard deviation

$$\sigma_{x_p} = \frac{\sigma}{y_p}\sqrt{\frac{pq}{n}} \qquad . \qquad . \qquad . \qquad . \qquad (20.1)$$

20.24. If the distribution in an indefinitely large sample be normal, we can take the values of y_p from the tables of the ordinate of the normal curve (Appendix Table 1). From tables carried to further places of decimals we have, for the various values of p which correspond to the deciles,.

				Value of y_p.
Median	.	.	.	. 0·3989423
Deciles 4 and 6	.	.	.	. 0·3863425
,, 3 and 7	.	.	.	. 0·3476926
,, 2 and 8	.	.	.	. 0·2799619
,, 1 and 9	.	.	.	. 0·1754983
Quartiles	.	.	.	. 0·3177766

Inserting these values of y_p in equation (20.1), we have the following values for the standard errors of the median, deciles, etc. :—

				Standard error is $\sigma/\sqrt{n}$ multiplied by
Median .	.	.	.	. 1·25331
Deciles 4 and 6	.	.	.	. 1·26804
,, 3 and 7	.	.	.	. 1·31800
,, 2 and 8	.	.	.	. 1·42877
,, 1 and 9	.	.	.	. 1·70942
Quartiles	.	.	.	. 1·36263

It will be seen that the influence of fluctuations of sampling on the several percentiles increases as we depart from the median : the standard error of the quartiles is nearly one-tenth greater than that of the median, and the standard error of the first or ninth decile more than one-third greater.

20.25. Consider further the influence of the form of the frequency-distribution on the standard error of the median, as this is an important form of average. For a distribution with a given number of observations and a given standard deviation the standard error varies inversely as y_p. Hence for a distribution in which y_p is small, for example a U-shaped distribution, the standard error of the median will be relatively high, and it will, in so far, be an undesirable form of average to employ. On the other hand, in the case of a distribution which has a high peak in the centre, so as to exhibit a value of y_p large compared with the standard deviation, the standard error of the median will be relatively low. We can create such a " peaked " distribution by superposing a normal curve with a small standard deviation on a normal curve with the same mean and a relatively large standard deviation. To give some idea of the reduction in the standard error of the median that may be effected by a

moderate change in the form of the distribution, let us find for what ratio of the standard deviations of two such curves, having the same area, the standard error of the median reduces to $\sigma/\sqrt{n}$, where σ is of course the standard deviation of the compound distribution.

Let σ_1, σ_2 be the standard deviations of the two distributions, and let there be $n/2$ observations in each. Then

$$\sigma = \sqrt{\frac{\sigma_1{}^2 + \sigma_2{}^2}{2}} \qquad . \qquad . \qquad . \qquad (20.2)$$

On the other hand, the value of y_p is

$$\left\{ \frac{1}{2\sqrt{2\pi}\sigma_1} + \frac{1}{2\sqrt{2\pi}\sigma_2} \right\} \sqrt{\frac{\sigma_1{}^2 + \sigma_2{}^2}{2}} \qquad . \qquad . \qquad (20.3)$$

Hence, the standard error of the median is

$$\sqrt{\frac{2\pi}{n}} \frac{\sigma_1\sigma_2}{\sigma_1 + \sigma_2} \qquad . \qquad . \qquad . \qquad (20.4)$$

(20.4) is equal to $\sigma/\sqrt{n}$ if

$$\frac{(\sigma_1 + \sigma_2)\sqrt{\sigma_1{}^2 + \sigma_2{}^2}}{2\sqrt{\pi}\sigma_1\sigma_2} = 1$$

and writing $\sigma_2/\sigma_1 = \rho$, that is if

$$\frac{(1 + \rho)\sqrt{1 + \rho^2}}{2\sqrt{\pi\rho}} = 1$$

or

$$\rho^4 + 2\rho^3 + (2 - 4\pi)\rho^2 + 2\rho + 1 = 0$$

This equation may be reduced to a quadratic and solved by taking $\rho + \dfrac{1}{\rho}$ as a new variable. The roots found give $\rho = 2\cdot2360 \ldots$ or $0\cdot4472 \ldots$, the one root being merely the reciprocal of the other. The standard error of the median will therefore be $\sigma/\sqrt{n}$, in such a compound distribution, if the standard deviation of the one normal curve is, in round numbers, about $2\frac{1}{4}$ times that of the other. If the ratio be greater, the standard error of the median will be less than $\sigma/\sqrt{n}$. The distribution for which the standard error of the median is exactly equal to $\sigma/\sqrt{n}$ is shown in fig. 20.1; it will be seen that it is by no means a very striking form of distribution; at a hasty glance it might almost be taken as normal. In the case of distributions of a form more or less similar to that shown, it is evident that we cannot at all safely estimate by eye alone the relative standard error of the median as compared with $\sigma/\sqrt{n}$.

20.26. In the case of a grouped frequency-distribution in which the number of observations is large enough to give a fairly smooth distribution, we can use an alternative form which does not involve a knowledge of the standard deviation of the distribution in a very large sample. In fact, in such a case the sample itself is large enough to give us a satisfactory

approximation to the distribution in an indefinitely large sample. Let f_p be the frequency per class-interval at the given percentile—simple inter-polation will give us the value with quite sufficient accuracy for practical purposes, and if the figures run irregularly they may be smoothed. Let σ be the value of the stan-dard deviation expressed in class-intervals, and let n be the number of obser-vations as before. Then, since y_p is the ordinate of the frequency-distribution when drawn with unit standard deviation and unit area, we must have

$$y_p = \frac{\sigma}{n} f_p$$

But this gives at once for the standard error *expressed in terms of the class-interval as unit*

FIG. 20.1.

$$\sigma_{x_p} = \frac{\sqrt{npq}}{f_p} \quad . \qquad . \qquad . \qquad . \qquad (20.5)$$

Example 20.4.—Consider the data of Table 6.7, page 94, giving the distribution of 8585 men according to height. Let us take these data to be a sample from the universe of men in the United Kingdom at that time. The number of observations is 8585, and the standard deviation 2·57 in., the distribution being approximately normal : $\sigma/\sqrt{n} = 0.027737$, and, multiplying by the factor 1·253 . . . given in the table in **20.24**, this gives 0·0348 as the standard error of the median, on the assumption of normality of the distribution.

Using the direct method of equation (20.5), we find the median to be 67·47 (7.20), which is very nearly at the centre of the interval with a frequency 1329. Taking this as being, with sufficient accuracy for our present purpose, the frequency per interval at the median, the standard error is

$$\tfrac{1}{2}\frac{\sqrt{8585}}{1329} = 0.0349$$

As we should expect, the value is practically the same as that obtained from the value of the standard deviation on the assumption of normality.

Three times the standard error is 0·1047, and we accordingly conclude that the median in the universe lies within about 0·1 inch of 67·47, the sample value, provided that the sampling is simple.

Example 20.5.—Let us find the standard error of the first and ninth deciles as another illustration. On the assumption that the distribution
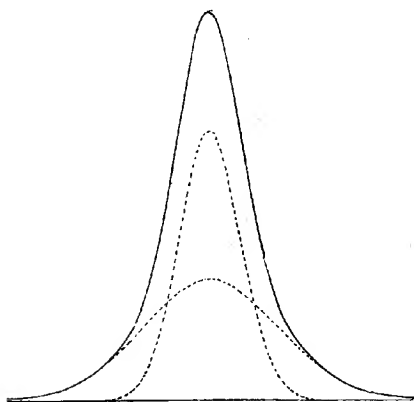
is normal, these standard errors are the same, and equal to 0.027787 $\times 1.70942 = 0.0474$. Using the direct method, we find by simple interpolation the approximate frequencies per interval at the first and ninth deciles respectively to be 590 and 570, giving standard errors of 0.0471 and 0.0488, mean 0.0479, slightly in excess of that found on the assumption that the frequency is given by the normal curve. The student should notice that the class-interval is, in this case, identical with the unit of measurement, and consequently the answer given by equation (20.5) does not require to be multiplied by the magnitude of the interval.

Correlation between Errors of Percentiles.

20.27. In finding the standard error of the difference between two percentiles in the same distribution, the student must be careful to note that the errors in two such percentiles are not independent. Consider the two percentiles for which the values of p and q are p_1 q_1, p_2 q_2, respectively, the first named being the lower of the two percentiles. These two percentiles divide the whole area of the frequency curve into three parts, the areas of which are proportional to q_1, $1 - q_1 - p_2$, and p_2. Further, since the errors in the first percentile are directly proportional to the errors in q_1, and the errors in the second percentile are directly proportional but of opposite sign to the errors in p_2, the correlation between errors in the two percentiles will be the same as the correlation between errors in q_1 and p_2, but of opposite sign. But if there be a deficiency of observations below the lower percentile, producing an error δ_1 in q_1, the missing observations will tend to be spread over the two other sections of the curve in proportion to their respective areas, and will therefore tend to produce an error

$$\delta_2 = -\frac{p_2}{p_1}\delta_1$$

in p_2. If, then, r be the correlation between errors in q_1 and p_2, ϵ_1 and ϵ_2 the respective standard errors, we have :

$$r\frac{\epsilon_2}{\epsilon_1} = -\frac{p_2}{p_1}$$

Or, inserting the values of the standard errors,

$$r = -\sqrt{\frac{p_2 q_1}{q_2 p_1}}$$

The correlation between the percentiles is the same in magnitude but opposite in sign; it is obviously positive, and consequently

$$\left.\begin{array}{c}\text{Correlation between errors} \\ \text{in two percentiles}\end{array}\right\} = +\sqrt{\frac{p_2 q_1}{q_2 p_1}} \quad . \quad . \quad (20.6)$$

If the two percentiles approach very close together, q_1 and q_2, p_1 and p_2 become sensibly equal to one another, and the correlation becomes unity, as we should expect.

Standard Error of Semi-interquartile Range.

20.28. Let us apply the above value of the correlation between percentiles to find the standard error of the semi-interquartile range for the normal curve. Inserting $q_1 = p_2 = \frac{1}{4}$, $q_2 = p_1 = \frac{3}{4}$, we find $r = \frac{1}{3}$. Hence the

25

standard error of the interquartile range is, applying the ordinary formula for the standard deviation of a difference, $2/\sqrt{3}$ times the standard error of either quartile, or the standard error of the *semi*-interquartile range $1/\sqrt{3}$ times the standard error of a quartile. Taking the value of the standard error of a quartile from the table in **20.24**, we have, finally,

$$\left.\begin{array}{l}\text{Standard error of the semi-}\\ \text{interquartile range in a}\\ \text{normal distribution}\end{array}\right\} = 0 \cdot 78672 \frac{\sigma}{\sqrt{n}} \qquad . \quad (20.7)$$

Of course the standard deviation of the interquartile, or semi-interquartile, range can readily be worked out in any particular case, using equation (20.5) and the value of the correlation given above; it is best to work out such standard errors from first principles, applying the usual formula for the standard deviation of the difference of two correlated variables (**16.2**).

20.29. If there is any failure of the conditions of simple sampling, the formulæ of the preceding sections cease, of course, to hold good. We need not, however, enter again into a discussion of the effect of removing the several restrictions, for the effect on the standard error of p was considered in detail in Chapter 19, and the standard error of any percentile is directly proportional to the standard error of p.

Standard Error of the Arithmetic Mean.

20.30. Let us now determine the standard error of the arithmetic mean.

Suppose we note separately at each drawing the value recorded on the first, second, third . . . and nth card of our sample. The standard deviation of the values on each separate card will tend in the long run to be the same, and identical with the standard deviation σ of x in an indefinitely large sample, drawn under the same conditions. Further, the value recorded on each card is (as we assume) uncorrelated with that on every other. The standard deviation of the sum of the values recorded on the n cards is therefore $\sqrt{n}\sigma$, and the standard deviation of the mean of the sample is consequently $1/n$th of this; or,

$$\sigma_m = \frac{\sigma}{\sqrt{n}} \qquad . \qquad . \qquad . \qquad (20.8)$$

This is a most important and frequently cited formula, and the student should note that it has been obtained without any reference to the size of the sample or to the form of the frequency-distribution. It is therefore of perfectly general application, if σ be known. We can verify it against our formula for the standard deviation of sampling in the case of attributes. The standard deviation of the number of successes in a sample of m observations is $\sqrt{mpq}$: the standard deviation of the total number of successes in n samples of m observations each is therefore $\sqrt{nmpq}$: dividing by n we have the standard deviation of the mean number of successes in the n samples, viz. $\sqrt{mpq}/\sqrt{n}$, agreeing with equation (20.8).

Example 20.6.—In the height distribution considered in Examples **20.4** and **20.5** we found that $\sigma/\sqrt{n} = 0 \cdot 0277$ approximately. This is then the standard error of the mean of the distribution.

If we regard the data as a simple sample from the universe of men in the United Kingdom, we may take the mean, *i.e.* 67·46 inches, as an estimate of the mean in the universe.   Three times the standard error is very small, 0·083 inch, and we can therefore locate the mean in the universe with considerable accuracy.

The standard error in this case, however, gives a misleading idea as to the accuracy attained in determining the average stature in the United Kingdom ; the sample was not chosen under conditions which gave every individual an equal chance of being chosen.

Comparison of the Standard Errors of the Median and the Mean.

20.31.   For a normal curve the standard error of the mean is to the standard error of the median approximately as 100 to 125 (*cf.* 20.24), and in general the standard errors of the two stand in a somewhat similar ratio for a distribution not differing largely from the normal form.   For the distribution of statures used as an illustration in Example 20.4, the standard error of the median was found to be 0·0349 ; the standard error of the mean is only 0·0277.   The distribution being very approximately normal, the ratio of the two standard errors, viz. 1·26, assumes almost exactly the theoretical magnitude.

As such cases as these seem on the whole to be more common and typical, we stated in 7.23 that the mean is *in general* less affected than the median by errors of sampling.   At the same time we also indicated the exceptional cases in which the median might be the more stable—cases in which the mean might, for example, be affected considerably by small groups of widely outlying observations, or in which the frequency-distribution assumed a form resembling fig. 20.1, but even more exaggerated as regards the height of the central " peak " and the relative length of the " tails."   Such distributions are not uncommon in some economic statistics, and they might be expected to characterise some forms of experimental error.   If, in these cases, the greater stability of the median is sufficiently marked to outweigh its disadvantages in other respects, the median may be the better form of average to use.   Fig. 20.1 represents a distribution in which the standard errors of the mean and of the median are the same.   Further, in some experimental cases it is conceivable that the median may be less affected by definite experimental errors, the average of which does not tend to be zero, than is the mean—this is, of course, a point quite distinct from that of errors of sampling.

Means of Two Samples.

20.32.   When we have two samples from some record which exhibit different means, a very common question which we wish to ask is : Can the difference be accounted for by sampling fluctuations, *i.e.* can the two samples have come from the same universe ?

If the two samples are independent and come from the same universe under simple conditions, evidently ϵ_{12}, the standard error of the difference of their means, is given by

$$\epsilon_{12}^2 = \sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right) \quad . \quad . \quad . \quad . \quad (20.9)$$

If an observed difference exceed three times the value of ϵ_{12} given by this formula, it can hardly be ascribed to fluctuations of sampling.   If, in

a practical case, the value of σ is not known *a priori*, we must substitute an observed value, and it would seem natural to take as this value the standard deviation in the two samples thrown together. If, however, the standard deviations of the two samples themselves differ more than can be accounted for on the basis of fluctuations of sampling alone (see below, **21.14**), we evidently cannot assume that both samples have been drawn from the same record : the one sample must have been drawn from a record or a universe exhibiting a greater standard deviation than the other. If two samples be drawn quite independently from different universes, indefinitely large samples from which exhibit the standard deviations σ_1 and σ_2, the standard error of the difference of their means will be given by

$$\epsilon_{12}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \qquad . \qquad . \qquad . \qquad . \qquad (20.10)$$

This is, indeed, the formula usually employed for testing the significance of the difference between two means in any case ; seeing that the standard error of the mean depends on the standard deviation only, and not on the mean, of the distribution, we can inquire whether the two universes from which samples have been drawn differ in mean *apart from any difference in dispersion.*

20.33. If two quite independent samples be drawn from the same universe, but instead of comparing the mean of the one with the mean of the other we compare the mean m_1 of the first with the mean m_0 of both samples together, the use of (20.9) or (20.10) is not justified, for errors in the mean of the one sample are correlated with errors in the mean of the two together. Following precisely the lines of the similar problem in **19.29**, we find that this correlation is $\sqrt{n_1/(n_1 + n_2)}$, and hence

$$\epsilon_{01}^2 = \sigma^2 \frac{n_2}{n_1(n_1 + n_2)} \qquad . \qquad . \qquad . \qquad . \qquad (20.11)$$

(For a complete treatment of this problem in the case of large samples drawn from two different universes, *cf.* ref. (463).)

Effect on Standard Error of Mean of Breakdown of Conditions for Simple Sampling.

20.34. Let us consider briefly the effect on the standard error of the mean if the conditions of simple sampling as laid down in **20.16** cease to apply.

If we do not draw from the same record all the time, but first draw a series of samples from one record, then another series from another record with a somewhat different mean and standard deviation, and so on, or if we draw the successive samples from essentially different parts of the same record, the standard error will be greatly increased.

For suppose we draw k_1 samples from the first record, for which the standard deviation (in an indefinitely large sample) is σ_1, and the mean differs by d_1 from the mean of all the records together (as ascertained by large samples in numbers proportionate to those now taken), k_2 samples from the second record, for which the standard deviation is σ_2, and the mean differs by d_2 from the mean of all the records together, and so on.

Then for the samples drawn from the first record the standard error of the mean will be $\sigma_1/\sqrt{n}$, but the distribution will centre round a value differing by d_1 from the mean for all the records together ; and so on for the samples drawn from the other records. Hence, if σ_m be the standard error of the mean in all the records taken together, N the total number of samples,

$$N\sigma_m{}^2 = S\left(k\frac{\sigma^2}{n}\right) + S(kd^2)$$

But the standard deviation σ_0 for all the records together is given by

$$N\sigma_0{}^2 = S(k\sigma^2) + S(kd^2)$$

Hence, writing $S(kd^2) = Ns_m{}^2$,

$$\sigma_m{}^2 = \frac{\sigma_0{}^2}{n} + \frac{n-1}{n}s_m{}^2 \qquad . \qquad . \qquad . \quad (20.12)$$

This equation corresponds precisely to equation (19.8), page 363. The standard error of the mean, if our samples are drawn from different records or from essentially different parts of the entire record, may be increased indefinitely as compared with the value it would have in the case of simple sampling. If, for example, we take the statures of samples of n men in a number of different districts of England, and the standard deviation of all the statures observed is σ_0, the standard deviation of the means for the different districts will not be $\sigma_0/\sqrt{n}$, but will have some greater value, dependent on the real variation in mean stature from district to district.

20.35. If we are drawing from the same record throughout, but always draw the first card from one part of that record, the second card from another part, and so on, and these parts differ more or less, the standard error of the mean will be decreased. For if, in large samples drawn from the subsidiary parts of the record from which the several cards are taken, the standard deviations are σ_1, σ_2, . . . σ_n, and the means differ by d_1, d_2, . . . d_n from the mean for a large sample from the entire record, we have :

$$\sigma_0{}^2 = \frac{1}{n}S(\sigma^2) + \frac{1}{n}S(d^2)$$

Hence,

$$\sigma_m{}^2 = \frac{1}{n^2}S(\sigma^2)$$

$$= \frac{\sigma_0{}^2}{n} - \frac{s_m{}^2}{n} \qquad . \qquad . \qquad . \quad (20.13)$$

The last equation again corresponds precisely with that given for the same departure from the rules of simple sampling in the case of attributes (equation (19.10), p. 365). If, to vary our previous illustration, we had measured the statures of men in each of n different districts, and then proceeded to form a set of samples by taking one man from each district for the first sample, one man from each district for the second sample, and so on, the standard deviation of the means of the samples so formed would be appreciably less than the standard error of simple

sampling $\sigma_0/\sqrt{n}$. As a limiting case, it is evident that if the men in each district were all of precisely the same stature, the means of all the samples so compounded would be identical; in such a case, in fact, $\sigma_0 = s_m$, and consequently $\sigma_m = 0$. To give another illustration, if the cards from which we were drawing samples had been arranged in order of the magnitude of X recorded on each, we would get a much more stable sample by drawing one card from each successive nth part of the record than by taking the sample according to our previous rules—*e.g.* shaking them up in a bag and taking out cards blindfold, or using some equivalent process.

The result is perhaps of some practical interest. It shows that, if we are actually taking samples from a large area, different districts of which exhibit markedly different means for the variable under consideration, and are limited to a sample of n observations, if we break up the whole area into n sub-districts, each as homogeneous as possible, and take a contribution to the sample from each, we will obtain a *more stable* mean by this orderly procedure than will be given, for the same number of observations, by any process of selecting the districts from which samples shall be taken by chance. There may, however, be a greater risk of biased error. These conclusions seem in accord with common sense.

20.36. Finally, suppose that, while our conditions (*a*) and (*b*) of **20.16** hold good, the magnitude of the variable recorded on one card drawn is no longer independent of the magnitude recorded on another card, *e.g.* that if the first card drawn at any sampling bears a high value, the next and following cards of the same sample are likely to bear high values also. In these circumstances, if r_{12} denote the correlation between the values on the first and second cards, and so on,

$$s_m{}^2 = \frac{\sigma^2}{n} + 2\frac{\sigma^2}{n^2}(r_{12} + r_{13} + \ldots + r_{23} + \ldots)$$

There are $n(n-1)/2$ correlations; and if, therefore, r is the arithmetic mean of them all, we may write:

$$\sigma_m{}^2 = \frac{\sigma^2}{n}[1 + r(n-1)] \qquad . \qquad . \qquad . \quad (20.14)$$

As the means and standard deviations of $x_1, x_2, \ldots x_n$ are all identical, r may more simply be regarded as the correlation coefficient for a table formed by taking all possible pairs of the n values in every sample. If this correlation be positive, the standard error of the mean will be increased, and for a given value of r the increase will be the greater, the greater the size of the samples. If r be negative, on the other hand, the standard error will be diminished. Equation (20.14) corresponds precisely to equation (19.12), page 366.

As was pointed out in **19.35**, the case when r is positive covers the case discussed in **20.34**; for if we draw successive samples from different records, such a positive correlation is at once introduced, although the drawings of the several cards *at each* sampling are quite independent of one another. Similarly, the case discussed in **20.35** is covered by the case of negative correlation, for if each card is always drawn from a separate and distinct part of the record, the correlation between any two x's will on the average be negative; if some one card be always drawn from a part

of the record containing low values of the variable, the others must on an average be drawn from parts containing relatively high values. It is as well, however, to keep the three cases distinct, since a positive or negative correlation may arise for reasons quite different from those considered in **20.34** and **20.35**.

SUMMARY.

1. A knowledge of the sampling distribution of a parameter enables us to ascertain the probability that a given sample will exhibit a value of the parameter between specified limits.

2. The sampling distribution of many parameters tends to the normal form, or at least a single-humped form, for large values of n, the number in the sample, if the sampling is simple.

3. This fact enables us to take a range of ± 3 times the standard error as providing limits within which a sample value of the parameter will probably lie ; with the further assumption of normality of the sampling distribution we can determine the probability that a sample value will lie within any specified limits.

4. In a large sample the values of parameters in the sample may be taken to be estimates of the values in the universe, if the sample is simple. Further, these values may be used instead of the values in the universe in calculating the standard errors of the parameters.

5. The standard error of the median of a normal distribution is given by

$$\text{s.e.} = 1 \cdot 25331 \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation in an indefinitely large sample and n is the number in the sample.

6. With the same notation the standard error of the arithmetic mean is

$$\text{s.e.} = \frac{\sigma}{\sqrt{n}}$$

whatever the form of the distribution.

7. If a series of samples of n is drawn from different universes or from different parts of a non-homogeneous universe,

$$\sigma_m{}^2 = \frac{\sigma_0{}^2}{n} + \frac{n-1}{n} s_m{}^2$$

where σ_m is the standard error of the mean, σ_0 is the standard deviation in all the samples taken together, and s_m is the standard deviation of means of indefinitely large samples about the mean of all samples.

8. If samples are drawn so that each member comes from a different section of a non-homogeneous universe,

$$\sigma_m{}^2 = \frac{\sigma_0{}^2}{n} - \frac{s_m{}^2}{n}$$

where σ_m, σ_0 and s_m are defined as before.

9. If there is a correlation between the results of the drawing of successive individuals,

$$\sigma_m{}^2 = \frac{\sigma^2}{n}[1 + r(n-1)]$$

where σ_m is the standard error of the mean, σ the standard deviation in an indefinitely large sample, and r is the mean correlation between the results of pairs of individuals.

EXERCISES.

20.1. If the sampling distribution of a parameter is normal, find the probability that a sample value will differ from the central value by more than twice the probable error.

20.2. In the height distribution of the United Kingdom given in Table 6.7, page 94, assumed to be normal, with mean 67·46 inches and standard deviation 2·57 inches, find the probability that an individual chosen in the same way as the members of the distribution will be between 5 and 6 feet in height.

20.3. For the data of the last column of Exercise 6.6, page 111, find the standard error of the median (154·7 lbs.) and the standard errors of the two quartiles (142·5 lbs. and 168·4 lbs.).

20.4. For the same distribution find the standard error of the semi-interquartile range.

20.5. The standard deviation of the same distribution is 21.3 lbs. Find the standard error of the mean and compare it with the standard error of the median (Exercise 20.3).

20.6. Taking the values of the median and the quartiles of the marriage distribution of Table 6.8, page 96, from Example 9.8, page 164, find their standard errors.

20.7. In the same distribution the mean is 29·4 years and the standard deviation 8 years, approximately. Find the standard error of the mean and compare it with that of the median.

20.8. For the same distribution find the standard error of the quartiles, assuming it to be normal with mean 29·4 years and standard deviation 8 years, and compare your results with those obtained in Exercise 20.6.

20.9. Find the standard error of the 27th percentile of the normal distribution.

20.10. (Imaginary data.) A random sample of 1000 men from the North of England shows their mean wage to be £2 7s. per week, with a standard deviation of £1 8s. A sample of 1500 men from the South of England gives a mean wage of £2 9s. per week, with a standard deviation of £2. Discuss the suggestion that the mean rate of wages varies as between the two regions.

20.11. Two universes have the same mean but the standard deviation of one is twice that of the other. Show that in samples of 500 from each drawn under simple random conditions the difference of the means will in all probability not exceed $0\cdot3\sigma$, where σ is the smaller standard deviation; and assuming the distribution of the difference of means to be normal, find the probability that it exceeds half that amount.

20.12. A random sample of 1000 farms in a certain year gives an average yield of wheat of 2000 lbs. per acre, with a standard deviation of 192 lbs. A random sample of 1000 farms in the following year gives an average yield of 2100 lbs. per acre, with a standard deviation of 224 lbs. Show that these data are consistent with the hypothesis that the average yields in the country as a whole were the same in the two years.

Would you modify this conclusion if the farms in the second sample were the same as those in the first?

20.13. Find the mean and median of the U-shaped distribution of Table 6.14, page 106, and compare their standard errors. (For the purpose of this exercise the median frequency may be found by simple interpolation, but this gives a value on the high side.)

20.14. The mean of a certain normal distribution is equal to the standard error of the mean of samples of 100 from that distribution. Find the probability that the mean of a sample of 25 from the distribution will be negative.

20.15. If it costs a shilling to draw one member of a sample, how much would it cost, in sampling from a universe with mean 100 and standard deviation 10, to take sufficient members to ensure that the mean of the sample in all probability would be within 0·01 per cent. of the true value? Find the extra cost necessary to double the precision.

20.16. Consider the data of Table 6.7, page 94, giving the distribution of men by height in each of the four countries which then formed part of the United Kingdom. The means and standard deviations of the four distributions are given in Exercise 7.1, page 131, and Exercise 8.1, page 152.

What is the standard error of the mean of a sample which consists of 400 men, 100 chosen at random from each of the four countries?

·THE SAMPLING OF VARIABLES—LARGE SAMPLES, CONTINUED.

The Problem.

21.1. We have just considered the standard errors of the most important measures of location, the median and the mean, and of certain measures of dispersion, the percentiles and the semi-interquartile range. We now proceed to discuss the standard errors of other important parameters, including the standard deviation, moments and correlation coefficients. All that we have said in regard to sampling distributions generally in **20.1** to **20.22** applies equally well to this chapter; and we shall throughout the following sections be thinking of simple sampling unless we state explicitly to the contrary.

Standard Errors of Moments.[1]

21.2. The data from which we calculate the moments are arranged into a certain number of groups. Suppose there are m such groups, and that the expected frequencies falling into them are $y_1, y_2, \ldots y_m$, where $y_1 + y_2 + \ldots + y_m = S(y) = n$, n being the number in the sample. The expected frequencies are, by definition, proportional to the frequencies in the various groups in a very large sample; and these, if the sampling is unbiased, are proportional to the frequencies in the various groups of the parent universe.

Let us in the first place recapitulate some of our earlier work by finding the standard error of one of the frequencies, say y_s, due to fluctuations of sampling.

The probability that an individual chosen from the universe falls into the sth group is $\dfrac{y_s}{n}$. The probability that it does not is $1 - \dfrac{y_s}{n}$. For n individuals the distribution of frequencies is given by the binomial

$$n\left\{\left(1 - \frac{y_s}{n}\right) + \frac{y_s}{n}\right\}^n$$

with an expected value y_s and a standard deviation

$$\sigma_{y_s} = \sqrt{n\,\frac{y_s}{n}\left(1 - \frac{y_s}{n}\right)}$$

Now, if the sample is large, we can take the observed frequency in the sth group in calculating the standard error of the frequency of that group.

[1] The student whose main interest lies in the practical application of the results of this chapter may prefer to omit paragraphs **21.2** to **21.8**.

Taking this observed frequency as our estimate of y_s, its standard error, σ_{y_s}, is given by

$$\sigma_{y_s}^2 = y_s\left(1 - \frac{y_s}{n}\right) \qquad . \qquad . \qquad . \qquad (21.1)$$

This, in another form, is our familiar result for the sampling of attributes.

21.3. We may now find the correlation between errors in y_s and errors in another group-frequency, say y_t. It is evident that such a correlation will exist, for if y_s falls below its expected value, some other frequencies must be increased.

We shall write a deviation of y_s as δy_s. (The symbol δ is not to be regarded as a number multiplying y_s, but is to be read together with y_s so that δy_s is a single symbol representing a single quantity.)

Since

$$S(y) = y_1 + y_2 + \ldots + y_m = n$$
$$S(\delta y) = \delta y_1 + \delta y_2 + \ldots + \delta y_m = 0$$

for the sum of deviations from the expected values must be zero.

We may now assume that, on the average, a deficiency δy_s in y_s will be spread over the remaining groups in proportion to the expected frequencies in those groups, *i.e.* that

$$\delta y_t = -\delta y_s \frac{y_t}{n - y_s}$$

Hence,

$$\delta y_s \delta y_t = -(\delta y_s)^2 \frac{y_t}{n - y_s} \qquad . \qquad . \qquad (21.2)$$

Now let us sum both sides of this equation for all values of the deviations δy_s and δy_t. By definition we shall get

$$\sigma_{y_s}\sigma_{y_t} r_{y_s y_t} = -\sigma_{y_s}^2 \frac{y_t}{n - y_s}$$

where $r_{y_s y_t}$ is the coefficient of correlation between δy_s and δy_t.

Hence, in virtue of (21.1),

$$\sigma_{y_s}\sigma_{y_t} r_{y_s y_t} = -\frac{y_s y_t}{n} \qquad . \qquad . \qquad . \qquad (21.3)$$

This is a more general case of the correlation between percentiles, which we considered in **20.27**.

Standard Error of the qth Moment about a Fixed Point.

21.4. By definition, the qth moment about an arbitrary point is μ_q', where

$$n\mu_q' = S(x_s^q y_s)$$

x being the variate *measured from the arbitrary point.*

Hence, writing as before, $\delta\mu_q'$ for the deviation in μ_q' due to deviations δy_s, we have:

$$n\delta\mu_q' = S(x_s^q \delta y_s)$$

Squaring both sides,

$$n^2(\delta\mu_q')^2 = (x_1{}^q\delta y_1 + x_2{}^q\delta y_2 + \ldots + x_n{}^q\delta y_n)^2$$
$$= S\{x_s{}^{2q}(\delta y_s)^2\} + 2S'(x_s{}^q x_t{}^q \delta y_s \delta y_t)$$

where S' denotes summation over all values of s and t except those for which $s = t$.

This equation holds for any one sample, and we have to sum it for all samples. Carrying out this summation first (in which s and t are fixed), and substituting from equations (21.1) and (21.3) on the right-hand side, we have:

$$n^2\sigma_{\mu_q'}^2 = S\left\{x_s{}^{2q}y_s\left(1 - \frac{y_s}{n}\right)\right\} - 2S'\left(\frac{x_s{}^q x_t{}^q y_s y_t}{n}\right)$$

$$= S(x_s{}^{2q}y_s) - \frac{1}{n}S(x_s{}^q y_s)S(x_t{}^q y_t)$$

$$= n\mu_{2q}' - n\mu_q'^2$$

Hence,

$$\sigma_{\mu_q'} = \sqrt{\frac{\mu_{2q}' - \mu_q'^2}{n}} \qquad . \qquad . \qquad . \qquad (21.4)$$

Example 21.1.—Let us find the standard error of the first moment, or mean h.

We have, from (21.4):

$$\sigma_{\mu_1'} = \sqrt{\frac{\mu_2' - \mu_1'^2}{n}}$$

$$= \sqrt{\frac{\mu_2' - h^2}{n}}$$

Now $\mu_2' - h^2$ is the second moment μ_2 about the mean, *i.e.* is σ^2. Hence,

$$\sigma_{\mu_1'} = \sigma_h = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

which is the result we have already found in **20.30**.

Correlation between Errors in the *q*th and *r*th Moments, both about the Same Fixed Point.

21.5. As in 21.4 we have:

$$n\delta\mu_q' = S(x_s{}^q \delta y_s)$$
$$n\delta\mu_r' = S(x_s{}^r \delta y_s)$$

Multiplying,

$$n^2\delta\mu_q'\delta\mu_r' = S(x_s{}^{q+r}\delta y_s{}^2) + S'\{(x_s{}^q x_t{}^r + x_s{}^r x_t{}^q)(\delta y_s \delta y_t)\}$$

and summing for all samples,

$$n^2\sigma_{\mu_q'}\sigma_{\mu_r'}r_{\mu_q'\mu_r'} = S(x_s{}^{q+r}\sigma_{y_s}^2) + S'\{(x_s{}^q x_t{}^r + x_s{}^r x_t{}^q)(\sigma_{y_s}\sigma_{y_t}r_{y_s y_t})\}$$

On substitution for $\sigma^2_{v_s}$ and $\sigma_{v_s}\sigma_{v_t}r_{v_s v_t}$ from (21.1) and (21.3), the right-hand side reduces to $n\mu'_{q+r} - n\mu'_q\mu'_r$, and hence,

$$\sigma_{\mu_{q'}}\sigma_{\mu_r}r_{\mu_{q'}\mu_{r'}} = \frac{\mu'_{q+r} - \mu'_q\mu'_r}{n} \qquad . \qquad . \qquad (21.5)$$

Standard Errors of the Moments about the Mean.

21.6. In **21.4** and **21.5** we have considered moments about a fixed point. In practice we have to deal more usually with moments about the mean *of the sample*. Since this mean is itself subject to sampling fluctuations, the standard errors of moments about the mean will not in general be the same as those about a fixed point.

If h is the mean we have, by definition,

$$n\mu_q = S\{(x_s - h)^q y_s\}$$
$$= S(x_s^q y_s) - qhS(x_s^{q-1}y_s) + T$$

where T is written generally for an expression involving h^2 and higher powers of h.

Now let h vary to $h + \delta h$, y_s vary to $y_s + \delta y_s$, and μ_q vary to $\mu_q + \delta\mu_q$. We have:

$$n(\mu_q + \delta\mu_q) = S\{x_s^q(y_s + \delta y_s)\} - q(h + \delta h)S\{x_s^{q-1}(y_s + \delta y_s)\} + T$$

Subtracting the equation for $n\mu_q$,

$$n\delta\mu_q = S(x_s^q\delta y_s) - q\delta hS(x_s^{q-1}y_s) - qS(x_s^{q-1}\delta h\delta y_s) + U$$
$$= n\delta\mu'_q - nq\mu'_{q-1}\delta h - nq\delta h\delta\mu'_{q-1} + U$$

where U will involve h and higher powers. We may neglect the term in $\delta h\delta\mu'_{q-1}$ as being small compared with the remaining terms. Squaring and summing for all samples,

$$\sigma^2_{\mu_q} = \sigma^2_{\mu_{q'}} + q^2\mu'^2_{q-1}\sigma^2_h - 2q\mu'_{q-1}\sigma_h\sigma_{\mu_{q'}}r_{h\mu_{q'}} + U$$

Substituting for $\sigma^2_{\mu_{q'}}$, etc., from (21.4) and (21.5),

$$\sigma^2_{\mu_q} = \frac{\mu'_{2q} - \mu'^2_q + q^2\mu'_2\mu'^2_{q-1} - 2q\mu'_{q-1}\mu'_{q+1}}{n} + U$$

Now put $h = 0$. U vanishes and the moments become moments about the mean and may therefore be written without dashes. Hence,

$$\sigma_{\mu_q} = \sqrt{\frac{\mu_{2q} - \mu_q^2 + q^2\mu_2\mu_{q-1}^2 - 2q\mu_{q-1}\mu_{q+1}}{n}} \qquad . \qquad . \qquad (21.6)$$

Correlation between Two Moments Both Measured about the Mean.

21.7. In a similar way it may be shown that

$$\sigma_{\mu_q}\sigma_{\mu_r}r_{\mu_q\mu_r} = \frac{\mu_{q+r} - \mu_q\mu_r + qr\mu_2\mu_{q-1}\mu_{r-1} - r\mu_{q+1}\mu_{r-1} - q\mu_{q-1}\mu_{r+1}}{n} \qquad (21.7)$$

We omit the algebra for the sake of brevity.

Correlation between Errors in a Moment about a Fixed Point and in a Moment about the Mean.

21.8. Let us first of all find the correlation between deviations in a group-frequency y_t and the moment μ_q' about a fixed point. We have:

$$n\mu_q' = S(x_s^q y_s)$$

Hence,

$$n\delta\mu_q'\delta y_t = \delta y_t S(x_s^q \delta y_s)$$
$$= x_t^q(\delta y_t)^2 + S'(x_s^q \delta y_s \delta y_t)$$

the summation S' being taken over all values of s except $s = t$.

Hence, summing for all samples,

$$n\sigma_{\mu_q'}\sigma_{y_t}r_{\mu_q'y_t} = x_t^q y_t\left(1 - \frac{y_t}{n}\right) - S'\left(\frac{x_s^q y_s y_t}{n}\right)$$
$$= y_t\left\{x_t^q - S\left(\frac{x_s^q y_s}{n}\right)\right\}$$
$$= y_t\{x_t^q - \mu_q'\}$$

Hence,

$$\sigma_{\mu_q'}\sigma_{y_t}r_{\mu_q'y_t} = \frac{y_t}{n}(x_t^q - \mu_q') \qquad . \qquad . \qquad . \quad (21.8)$$

Similarly, for the product-sum of deviations in y_t and the moment μ_q about the mean, we have:

$$\sigma_{\mu_q}\sigma_{y_t}r_{\mu_q y_t} = \frac{y_t}{n}(x_t^q - \mu_q') - \frac{qy_t}{n}(x_t - h)\mu_{q-1}'$$

$$+ \text{terms in } h \text{ and higher powers}$$

Putting $h = 0$, the right-hand side reduces to

$$\frac{y_t}{n}(x_t^q - \mu_q - qx_t\mu_{q-1}) \quad . \qquad . \qquad . \qquad . \quad (21.9)$$

For the product-sum of errors in μ_q' and μ_r,

$$n\delta\mu_q' = S(x_s^q \delta y_s)$$
$$\delta\mu_r = \delta\mu_r' - r\delta h\mu_{r-1}' + U$$

where U, as before, denotes an expression involving h and higher powers.

Hence,

$$n\delta\mu_q'\delta\mu_r = S(x_s^q \delta y_s \delta\mu_r') - S(x_s^q \delta y_s r\delta h\mu_{r-1}') + U$$

Summing for all deviations,

$$\sigma_{\mu_q'}\sigma_{\mu_r}r_{\mu_q'\mu_r} = S(x_s^q \sigma_{y_s}\sigma_{\mu_r}r_{y_s\mu_r'}) - S(x_s^q r\mu_{r-1}'\sigma_{y_s}\sigma_h r_{y_s h}) + U$$

and substituting from (21.8) and (21.9) the right-hand side becomes

$$\frac{\mu_{q+r}' - \mu_q'\mu_r'}{n} - \frac{r\mu_{q+1}'\mu_{r-1}'}{n} + U$$

Put $h = 0$. Then,

$$\sigma_{\mu_q'}\sigma_{\mu_r}r_{\mu_q'\mu_r} = \frac{\mu_{q+r} - \mu_q\mu_r - r\mu_{q+1}\mu_{r-1}}{n} \qquad . \quad . \quad . \quad (21.10)$$

Use of Sheppard's Corrections in Evaluating Standard Errors.

21.9. Theoretically, Sheppard's corrections for grouping are not to be used in evaluating the moments which enter into the general equations for standard errors obtained in the previous sections. For, as the corrected values differ from the uncorrected values only by constants depending on the width of the interval, the sampling deviations of corrected and uncorrected moments are equal, and hence so are their standard errors. But the standard errors of uncorrected moments are given by the equations we have obtained in the foregoing section, and hence those equations are applicable to corrected moments provided that the uncorrected values are used in them.

In practice, however, it seems to make very little difference which moments we use, unless the sample is very large indeed. But as the uncorrected values have to be obtained before the corrected values can be calculated, and are therefore usually available, it is as well to use the uncorrected values wherever possible.

Standard Error of the Variance.

21.10. Armed with the general results of the foregoing sections, the methods of which are due to Karl Pearson (ref. (460)), we can discuss the standard errors of a large class of parameters.

From equation (21.6), putting $q = 2$, we have, since $\mu_1 = 0$,

$$\sigma_{\mu_2} = \sqrt{\frac{\mu_4 - \mu_2^2}{n}} \qquad . \qquad . \qquad . \qquad (21.11)$$

which gives the standard error of the variance μ_2.

If the parent universe is normal,

$$\mu_2 = \sigma^2, \qquad \mu_4 = 3\sigma^4 \quad \textbf{(10.23)}$$

and hence,

$$\sigma_{\mu_2} = \sqrt{\frac{3\sigma^4 - \sigma^4}{n}} = \sigma^2\sqrt{\frac{2}{n}}$$

$$= \mu_2\sqrt{\frac{2}{n}} \qquad . \qquad . \qquad . \qquad . \qquad (21.12)$$

Standard Error of the Standard Deviation.

21.11. If μ_2 is the variance, we have:

$$\mu_2 = \sigma^2$$

Hence,

$$\mu_2 + \delta\mu_2 = (\sigma + \delta\sigma)^2$$
$$= \sigma^2 + 2\sigma\delta\sigma + (\delta\sigma)^2 \quad .$$

Neglecting $\delta\sigma^2$ in comparison with $\delta\sigma$,

$$\delta\mu_2 = 2\sigma\delta\sigma$$

Squaring and summing for all samples,

$$\sigma_{\mu_2}^2 = 4\sigma^2\sigma_\sigma^2$$

Hence,

$$\sigma_\sigma = \frac{1}{2\sigma}\sigma_{\mu_2} = \sqrt{\frac{\mu_4 - \mu_2{}^2}{4\mu_2 n}} \qquad . \qquad . \qquad . \quad (21.13)$$

If the parent distribution is normal this reduces to

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2n}} \qquad . \qquad . \qquad . \quad (21.14)$$

21.12. The form of equation (21.14) has been widely used for the standard error of σ without due regard to the nature of the parent universe, and the student should guard against this mistake.

We have, in fact, from (21.13):

$$\sigma_\sigma = \frac{\sqrt{\mu_2}}{\sqrt{2n}}\sqrt{\frac{1}{2}\left(\frac{\mu_4}{\mu_2{}^2} - 1\right)}$$

$$= \frac{\sigma}{\sqrt{2n}}\left(1 + \frac{\beta_2 - 3}{2}\right)^{\frac{1}{2}}$$

How far σ_σ can be taken to be the value (21.14) therefore depends on how close the factor $\left(1 + \frac{\beta_2 - 3}{2}\right)^{\frac{1}{2}}$ is to unity, *i.e.* depends on the kurtosis of the parent distribution.

The following table shows the value of this factor for various values of β_2 :—

β_2	$\left(1 + \frac{\beta_2 - 3}{2}\right)^{\frac{1}{2}}$
2	0·7071
3	1·0000
4	1·2247
5	1·4142
6	1·5811
7	1·7321
8	1·8708
9	2·0000

It thus appears that if the universe is leptokurtic the real standard error is greater than that given by the assumption of normality, and may be twice as great or even more. If the universe is platykurtic the real standard error is less than the " normal " value.

If $\frac{\beta_2 - 3}{2}$ is small, the factor $\left(1 + \frac{\beta_2 - 3}{2}\right)^{\frac{1}{2}}$ is approximately $1 + \frac{\beta_2 - 3}{4}$. This differs from unity by more than 5 per cent. if β_2 is less than 2·8 or more than 3·2. Hence, values of β_2 lying outside the range 2·8 to 3·2 (and they are more common than not in practice) will give an error of more than 5 per cent. if the universe is assumed to be normal.

Example 21.2.—For the height distribution of Table 6.7, page 94, we have found that $\sigma = 2·57$ inches, $n = 8585$. The universe may be taken to

be normal, for β_2 from the sample is 3·149 (Example 9.9, page 165) and hence the standard error of $\sigma = \dfrac{2\cdot57}{\sqrt{2 \times 8585}} = 0\cdot02$ approximately.

Hence, we may say that the s.d. in the universe almost certainly lies in the range $2\cdot57 \pm 0\cdot06$, assuming that the sampling is simple.

Example 21.3.—The distribution of Australian marriages of Table 6.8, page 96, has uncorrected moments μ_2 and μ_4, in class-intervals, as follows:

$$\mu_2 = 7\cdot0570$$
$$\mu_4 = 408\cdot7382 \qquad \text{(Example 9.2, page 159.)}$$

Hence,

$$\sigma = \sqrt{\mu_2} = 2\cdot6565$$

The standard error of $\sigma = \sqrt{\dfrac{\mu_4 - \mu_2{}^2}{4\mu_2 n}}$

$$= \sqrt{\frac{408\cdot7382 - (7\cdot0570)^2}{4 \times 7\cdot0570 \times 301,785}}$$

$$= 0\cdot00649 \text{ class-intervals}$$

As we should expect from such a large sample, the standard error is very small, and we conclude that the standard deviation of the parent lies in the range $2\cdot6565 \pm 0\cdot0195$.

It may be pointed out that if we take these data as a sample of Australian marriages in general, we may be violating the conditions of simple sampling, for the distribution most likely changes from year to year.

Example 21.4.—In the previous example we worked throughout with uncorrected values. The corrected moments (Example 9.4, page 160) are:

$$\mu_2 = 6\cdot9736$$
$$\mu_4 = 405\cdot2389$$

We then have, for the corrected value of σ,

$$\sigma = \sqrt{6\cdot9736}$$
$$= 2\cdot641$$

But the standard error of σ is 0·00649 as in the previous example, for we must use the uncorrected values in calculating it.

As a matter of fact, if we had used the corrected values we should have found the value 0·00651—a practically negligible difference even for a sample of this size.

Finally, let us compare this value with that given by the assumption of normality. We have:

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2n}} = \frac{2\cdot6565}{\sqrt{603,570}}$$

$$= 0\cdot00342 \text{ class-intervals}$$

26

i.e. only about half the true value. This is in accordance with the table of page 400, for β_2 is over 8.

Comparative Effects of Sampling Fluctuations and Corrections for Grouping.

21.13. Writing temporarily σ_1^2 for the uncorrected value of the variance and σ_2^2 for the corrected value, we have :

$$\sigma_2^2 = \sigma_1^2 - \frac{h^2}{12}$$

or

$$\frac{\sigma_2^2}{\sigma_1^2} = 1 - \frac{1}{12}\frac{h^2}{\sigma_1^2}$$

If the class-interval is chosen so as to make the number of intervals d, then $6\sigma_1$ would be about dh and $\dfrac{h}{\sigma_1}$ about $\dfrac{6}{d}$.   Hence,

$$\frac{\sigma_2^2}{\sigma_1^2} = 1 - \frac{3}{d^2}$$

or, since $\dfrac{3}{d^2}$ is small,

$$\frac{\sigma_2}{\sigma_1} = 1 - \frac{3}{2d^2}$$

For instance, if d is 20, the corrected value is about 0·375 per cent. less than the uncorrected value.

Now, for a normal universe,

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2n}}$$

and if n is, say, 1000, the standard error is $\dfrac{\sigma}{44\cdot72} = 0\cdot0224\sigma = 2\cdot24$ per cent. of σ. Thus Sheppard's correction amounts to no more than about one-sixth of the standard error, and to make it gives an almost misleading idea of precision in most practical cases.

It was for this reason that we recommended (**8.11** and **11.29**) that the Sheppard corrections should not be applied if the total frequency is less than 1000.   On the other hand, in Examples 21.3 and 21.4 the correction is large compared with the standard error and can reasonably be made, owing to the largeness of the sample.

Comparison of Standard Deviations of Two Samples.

21.14.   As in 20.32, where we considered the comparison of the means of two samples, if the samples are independent and come from the same universe the standard error of the difference of their standard deviations is given by

$$\epsilon_{12}^2 = \frac{\mu_4 - \mu_2^2}{4\mu_2}\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\} \qquad . \qquad . \qquad . \quad (21.15)$$

where n_1, n_2 are the numbers in the samples, or, if the universe be normal,

$$\epsilon_{12} = \frac{r^2}{2}\left\{\frac{1}{n_1}+\frac{1}{n_2}\right\} \qquad . \qquad . \qquad . \qquad (21.16)$$

If the two samples are drawn from different universes with constants μ_2, μ_4 and ν_2, ν_4, the standard error of the difference of the standard deviations is given by

$$\epsilon_{12}^2 = \frac{\mu_4 - \mu_2{}^2}{4\mu_2 n_1} + \frac{\nu_4 - \nu_2{}^2}{4\nu_2 n_2} \qquad . \qquad . \qquad (21.17)$$

or

$$\epsilon_{12}^2 = \frac{\sigma_1{}^2}{2n_1} + \frac{\sigma_2{}^2}{2n_2} \qquad . \qquad . \qquad . \qquad (21.18)$$

if the universe be normal.

Again, if the standard deviation of one sample is compared with the standard deviation of the two samples when pooled, the standard error of the difference is, if the distribution be normal,

$$\epsilon_{01}^2 = \frac{\sigma^2}{2}\,\frac{n_2}{n_1(n_1 + n_2)} \qquad . \qquad . \qquad . \qquad (21.19)$$

These results can be used to test the significance of differences between standard deviations precisely as the equations of **20.32** and **20.33** were used to test the significance of differences between means.

Standard Error of Third and Fourth Moments about the Mean.

21.15. From equation (21.6), putting $q = 3$,

$$\sigma_{\mu_3} = \sqrt{\frac{\mu_6 - \mu_3{}^2 - 6\mu_4\mu_2 + 9\mu_2{}^3}{n}} \qquad . \qquad . \qquad (21.20)$$

If the distribution is normal,

Hence,
$$\mu_6 = 15\sigma^6, \qquad \mu_4 = 3\sigma^4, \qquad \mu_3 = 0, \qquad \mu_2 = \sigma^2$$

$$\sigma_{\mu_3} = \frac{\sigma^3}{\sqrt{n}}\sqrt{15 - 18 + 9} = \sigma^3\sqrt{\frac{6}{n}} \qquad . \qquad . \qquad (21.21)$$

Similarly, from equation (21.6), putting $q = 4$,

$$\sigma_{\mu_4} = \sqrt{\frac{\mu_8 - \mu_4{}^2 - 8\mu_5\mu_3 + 16\mu_2\mu_3{}^2}{n}} \qquad . \qquad . \qquad (21.22)$$

If the distribution is normal, $\mu_8 = 105\sigma^8$, $\mu_5 = 0$.
Hence,

$$\sigma_{\mu_4} = \frac{\sigma^4}{\sqrt{n}}\sqrt{105 - 9}$$

$$= \sigma^4\sqrt{\frac{96}{n}} \qquad . \qquad . \qquad . \qquad (21.23)$$

Example 21.5.—For the height distribution of Table 6.7 we have (Example 9.1, page 156):

$$\mu_2 \text{ (uncorrected)} = 6\cdot6168$$
$$\mu_3 \text{ (uncorrected)} = -0\cdot2078$$
$$\mu_4 \text{ (uncorrected)} = 137\cdot6892$$

and from Example 9.3, page 160:

$$\mu_2 \text{ (corrected)} = 6\cdot5335$$
$$\mu_3 \text{ (corrected)} = -0\cdot2078$$
$$\mu_4 \text{ (corrected)} = 134\cdot4100$$

We did not calculate higher moments, and hence cannot use equations (21.20) and (21.22) with these data. The distribution is, however, approximately normal. Hence, from (21.21),

$$\sigma_{\mu_3} = \sigma^3 \sqrt{\frac{6}{8585}} = 0\cdot45 \text{ approximately}$$

The value of μ_3 cannot therefore be judged significantly different from zero, which is what we should expect, for we have assumed the universe to be normal.

From (21.23) we have:

$$\sigma_{\mu_4} = \sigma^4 \sqrt{\frac{96}{8585}}$$

$$= 4\cdot63 \text{ approximately}$$

These are calculated from the uncorrected value of σ. We may infer that μ_4 (corrected) lies within the range $134\cdot41 \pm 13\cdot89$. The Sheppard correction is only $3\cdot28$, and is submerged in the possible sampling deviation, even for a sample of 8585. What we have said in **21.13** applies, in fact, *a fortiori* to the higher moments.

21.16. It will be evident that the standard errors of moments of high order are very large; for the moments increase rapidly, and the standard error of the moment of order q depends on the moment of order $2q$. For example, in the normal distribution, for $q = 6$, $\mu_{2q} = 10,395\sigma^{12}$ and σ_{μ_6} will be of the order $\dfrac{100\sigma^6}{\sqrt{n}}$, whereas $\mu_6 = 15\sigma^6$. Unless, therefore, n is at least 400, the range $3\sigma_{\mu_6}$ will be greater than the value of μ_6, and hence we cannot locate the value of μ_6 in the universe with any exactness. Our approximations, in fact, break down if the deviations are large.

The large sampling errors of moments of high orders prevent the use of moments higher than the fourth in most practical problems.

Correlation between Errors in Mean and Standard Deviation.

21.17. From equation (21.10), putting $q = 1$, $r = 2$, and remembering that $\mu_1 = 0$, we have:

$$\frac{\sigma}{\sqrt{n}} \sigma_{\mu_2} r_{h\mu_2} = \frac{\mu_3}{n}$$

Hence, if $\mu_3 = 0$, errors in the mean and variance, and hence in the mean and s.d., are uncorrelated. In particular, we have the important result that errors in the mean and s.d. in a normal universe are uncorrelated.

Standard Error of the Coefficient of Variation.

21.18. The coefficient of variation V is defined as

$$V = \frac{100\sigma}{h}$$

$$= \frac{100\sqrt{\mu_2}}{h}$$

Hence,

$$V + \delta V = \frac{100\sqrt{\mu_2 + \delta\mu_2}}{h + \delta h}$$

$$= \frac{100\sqrt{\mu_2}}{h}\left(1 + \frac{\delta\mu_2}{\mu_2}\right)^{\frac{1}{2}}\left(1 + \frac{\delta h}{h}\right)^{-1}$$

$$= V\left\{1 + \frac{\delta\mu_2}{2\mu_2}\right\}\left\{1 - \frac{\delta h}{h}\right\}$$

Neglecting quantities small compared with $\delta\mu_2$ and δh, this becomes

$$V\left\{1 + \frac{\delta\mu_2}{2\mu_2} - \frac{\delta h}{h}\right\}$$

Hence,

$$\frac{\delta V}{V} = \frac{\delta\mu_2}{2\mu_2} - \frac{\delta h}{h}$$

$$\frac{(\delta V)^2}{V^2} = \frac{(\delta\mu_2)^2}{4\mu_2^2} + \frac{(\delta h)^2}{h^2} - \frac{1}{\mu_2 h}\delta\mu_2\delta h$$

Summing for all samples we have :

$$\frac{\sigma_V^2}{V^2} = \frac{\sigma_{\mu_2}^2}{4\mu_2^2} + \frac{\sigma_h^2}{h^2} - \frac{1}{\mu_2 h}\sigma_{\mu_2}\sigma_h r_{\mu_2 h}$$

If the distribution is normal :

$$\sigma_{\mu_2}^2 = \frac{2\sigma^4}{n}, \qquad \sigma_h^2 = \frac{\sigma^2}{n}$$

and $r_{\mu_2 h} = 0$ **(21.17)**.

Hence,

$$\frac{\sigma_V^2}{V^2} = \frac{1}{2n} + \frac{\sigma^2}{h^2 n}$$

$$= \frac{1}{2n}\left\{1 + \frac{2V^2}{10^4}\right\}$$

Hence,

$$\sigma_V = \frac{V}{\sqrt{2n}}\sqrt{1 + \frac{2V^2}{10^4}} \qquad . \qquad . \qquad . \quad (21.24)$$

In many practical cases the second term differs little from unity and $\dfrac{V}{\sqrt{2n}}$ will give a sufficiently precise result.

Standard Error of β_1 and β_2.

21.19. The standard errors of β_1 and β_2 can be deduced in a similar manner.

In fact,

$$\beta_1 = \frac{\mu_3{}^2}{\mu_2{}^3}$$

$$\beta_1 + \delta\beta_1 = \frac{(\mu_3 + \delta\mu_3)^2}{(\mu_2 + \delta\mu_2)^3}$$

which, after some reduction, gives

$$\delta\beta_1 = \frac{2\mu_3\delta\mu_3}{\mu_2{}^3} - \frac{3\mu_3{}^2}{\mu_2{}^4}\delta\mu_2$$

Squaring and summing for all samples :

$$\sigma_{\beta_1}^2 = \frac{4\mu_3{}^2}{\mu_2{}^6}\sigma_{\mu_3}^2 + \frac{9\mu_3{}^4}{\mu_2{}^8}\sigma_{\mu_2}^2 - \frac{12\mu_3{}^3}{\mu_2{}^7}\sigma_{\mu_3}\sigma_{\mu_2}r_{\mu_3\mu_2}$$

$$n\sigma_{\beta_1}^2 = \frac{4\mu_3{}^2}{\mu_2{}^6}(\mu_6 - \mu_3{}^2 - 6\mu_4\mu_2 + 9\mu_2{}^3)$$

$$+ \frac{9\mu_3{}^4}{\mu_2{}^8}(\mu_4 - \mu_2{}^2) - \frac{12\mu_3{}^3}{\mu_2{}^7}(\mu_5 - 4\mu_2\mu_3)$$

In terms of β_1, β_2, β_3 and β_4 (see p. 161, footnote, for definition of the higher β's),

$$\sigma_{\beta_1}^2 = \frac{\beta_1}{n}\{4\beta_4 - 24\beta_2 + 36 + 9\beta_1\beta_2 - 12\beta_3 + 35\beta_1\} \qquad . \quad (21.25)$$

Similarly,

$$\sigma_{\beta_2}^2 = \frac{1}{n}\{\beta_6 - 4\beta_2\beta_4 + 4\beta_2{}^3 - \beta_2{}^2 + 16\beta_2\beta_1 - 8\beta_3 + 16\beta_1\} \quad . \quad (21.26)$$

The labour of evaluating these quantities may be obviated by the use of tables given in " *Tables for Statisticians and Biometricians, Part I.* "

21.20. There is here one important point to be noted. In equation (21.24), if $V = 0$, $\sigma_V = 0$. Similarly, in equation (21.25), if $\beta_1 = 0$, $\sigma_{\beta_1} = 0$. It might be thought from this that if in a large sample we find in the one case that $V = 0$ (and hence that $\sigma = 0$), or in the other case that the distribution is symmetrical, then $V = 0$ or $\beta_1 = 0$ in the universe. This is not necessarily true.

V will vanish only if all members of the sample give the same value of the variate. If the sample is large, it will be evident that if there is any variation in the parent it must be small ; but it is not impossible that members should exist showing deviations from the observed value. The explanation is to be found in the terms which we have neglected in our approximations. These, though in general small compared with the terms retained, may be important if the terms retained themselves

vanish. Furthermore, our assumption that the sample value may be assumed to be the parent value may be unjustified if both are very small compared with their difference. Equations such as (21.24) and (21.25) must, therefore, be treated carefully in the neighbourhood of values which cause them to vanish.

21.21. From the foregoing work the student will have no difficulty in accepting the statement that it is possible to calculate the standard error of any quantity which is expressible as a function of the moments. Such a standard error would, however, be applicable only to a value which had actually been calculated from the moments, and not arrived at by some other means. We shall not pursue the subject further in this book, but we may point out that the standard errors of certain quantities, such as an approximation to the Pearson measure of skewness (**9.12**), have been tabulated in " *Tables for Statisticians and Biometricians* " for different values of β_1 and β_2. The same tables also contain some results of interest in connection with the sampling distributions of range.

We now turn to the parameters of multivariate universes, the correlation coefficients, regression coefficients, and some of the measures of association.

Standard Error of the Correlation Coefficient.

21.22. For samples from a normal universe the standard error of the correlation coefficient is given by

$$\sigma_r = \frac{1 - r^2}{\sqrt{n}} \qquad . \qquad . \qquad . \qquad (21.27)$$

A proof of this result would take us beyond the scope of the present work. The student who is acquainted with the differential and integral calculus may refer to ref. (459).

The formula applies also to partial correlations.

21.23. Formula (21.27) is sometimes used to estimate the precision of correlation coefficients obtained by the use of the product-moment formula without reference to the nature of the universe. This practice is hardly to be commended, although sometimes there is nothing better to do. It is, however, possible to generalise the procedure of sections **21.2** to **21.8** to the bivariate case, and it may be shown that

$$\frac{\sigma_r^2}{r^2} = \frac{1}{n}\left\{ \frac{\mu_{22}}{\mu_{11}^2} + \frac{1}{4}\frac{\mu_{40}}{\mu_{20}^2} + \frac{1}{4}\frac{\mu_{04}}{\mu_{02}^2} + \frac{1}{2}\frac{\mu_{22}}{\mu_{20}\mu_{02}} - \frac{\mu_{31}}{\mu_{11}\mu_{20}} - \frac{\mu_{13}}{\mu_{11}\mu_{02}} \right\} \quad . \quad (21.28)$$

(For the definition of the bivariate moments, see footnote, p. 214.)

In addition, if the regression is linear, denoting the β_2's of the two variates considered separately by β_2, β_2',

$$\sigma_r^2 = \frac{(1 - r^2)^2}{n}\left\{ 1 - \frac{r^2}{4(1 - r^2)}(\beta_2 - 3 + \beta_2' - 3) \right\} \qquad . \quad (21.29)$$

which reduces to (21.27) if the kurtosis is zero.

If the distribution is not normal and r is not small, the difference between the values given by (21.27) and (21.29) may be considerable ; but it may be noticed that the value given by (21.27) is less than that given by (21.29)

if the distribution is platykurtic for both variates, and greater if the distribution is leptokurtic for both variates.

21.24.   In particular, it may be shown that for a 2×2 table in which the frequencies are (AB), $(A\beta)$, (αB) and $(\alpha\beta)$, the standard error of the correlation coefficient calculated by the product-moment method on the assumption that the frequencies are concentrated at points is given by

$$\sigma_r{}^2 = \frac{1}{n}\left\{ 1 - r^2 + (r + \tfrac{1}{2}r^3)\frac{[(A) - (\alpha)][(B) - (\beta)]}{\sqrt{(A)(\alpha)(B)(\beta)}} \right.$$
$$\left. - \tfrac{3}{4}r^2\left[\frac{[(A) - (\alpha)]^2}{(A)(\alpha)} + \frac{[(B) - (\beta)]^2}{(B)(\beta)}\right]\right\} \quad (21.30)$$

21.25.   The standard error of tetrachoric r, as calculated in the manner of **13.23**, is given by very complicated expressions which we do not reproduce.   The student may be referred to ref. (465) for an approximate form and certain tables to facilitate the arithmetic.

Example 21.6.—In the data of Table 11.3, page 199, we found that the correlation between the stature of the father and the stature of the son was 0·51.   Regarding these data as a sample of 1078 from the universe of fathers and sons, we have :

$$\text{Standard error of } r = \frac{1 - r^2}{\sqrt{n}} = \frac{1 - (0·51)^2}{\sqrt{1078}}.$$
$$= 0·023 \text{ approximately}$$

Hence, if the sampling was simple, the correlation in the universe most probably lies within 0·44 and 0·58.   It is thus undoubtedly real.

Example 21.7.—In considering data from 14,416 cows, J. F. Tocher found a negative correlation of 0·0796 between yield of milk per week and percentage of butter fat.   Is this significant, *i.e.* could it have arisen from an uncorrelated universe by sampling fluctuations ?

If $r = 0$,

$$\sigma_r = \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{14,416}}$$
$$= 0·008$$

The correlation observed is ten times this, and small though it is, could not have arisen from sampling fluctuations.

In this example we may reiterate the caution to be observed in inferring from the sample anything about the universe (cows in the United Kingdom) as a whole.   The records were, in fact, taken by the Scottish Milk Records Association from constituent associations at various years between 1908 and 1923.   The conditions of simple sampling may, therefore, have been violated both in regard to time and in regard to place.

Standard Error of the Coefficient of Regression.

21.26.   The standard error of the coefficient of regression from a normal universe is given by

$$\sigma_{b_{12}} = \frac{\sigma_1 \sqrt{1 - r_{12}^2}}{\sigma_2 \sqrt{n}} = \frac{\sigma_{1.2}}{\sigma_2 \sqrt{n}} \qquad . \qquad . \qquad . \quad (21.31)$$

This again applies to a regression coefficient of any order, total or partial, *i.e.* in terms of our general notation, k denoting any collection of secondary subscripts other than 1 or 2,

$$\left. \begin{array}{c} \text{Standard error of } b_{12.k} \\ \text{for a normal distribution} \end{array} \right\} = \frac{\sigma_{1.2k}}{\sigma_{2.k} \sqrt{n}}$$

The Correlation Ratio and Coefficient of Multiple Correlation.

21.27. It has been shown that the sampling distributions of the correlation ratio and the multiple correlation coefficient from normal universes do *not* tend to the normal form for large samples, although they do give single-humped distributions. The use of a standard error in such cases must be made with great caution, and it is probably better to apply one of the tests of significance which we shall consider later in connection with the theory of small samples. The formula usually given for the standard error of the correlation ratio is an approximate one :

$$\sigma_\eta = \frac{1 - \eta^2}{\sqrt{n}} \qquad . \qquad . \qquad . \qquad . \quad (21.32)$$

21.28. Somewhat similar remarks apply to the coefficient $\zeta = \eta^2 - r^2$ which, as we saw in **13.8**, may be used to test the linearity of regression. The use of a standard error for ζ in an attempt to gauge the significance of a departure from linearity has been subjected to very damaging criticism by R. A. Fisher.

Example 21.8.—Consider the data of Example 14.2, page 272 (relation between pauperism, age of population and number of population).
We found :

$$x_1 = 0.325 x_2 + 1.383 x_3 - 0.383 x_4$$

Taking this to be given by a random sample from a normal universe, is the value 0.325 significant ?
We have :

$$\sigma_{b_{12.34}} = \frac{\sigma_{1.234}}{\sigma_{2.34} \sqrt{n}} = \frac{\sigma_{1.234} \sqrt{1 - r_{21.34}^2}}{\sigma_{2.134} \sqrt{n}}$$

$$= \frac{22.8 \sqrt{1 - 0.457^2}}{32.1 \sqrt{32}}$$

$$= 0.11$$

The coefficient $b_{12.34}$ is therefore significant.
In this example the number in the sample is not as large as one might wish and the standard error is probably underestimated ; but if any doubt exists it is possible to make more definite tests by the methods of Chapter 23.

Standard Error of Coefficient of Association.

21.29. We may refer briefly to the quantities treated in Chapters 3, 4 and 5 in considering the association of attributes.

The coefficient of association, Q, defined in **3.15**, has a standard error given by

$$\sigma_Q = \frac{1-Q^2}{2}\sqrt{\frac{1}{(AB)}+\frac{1}{(A\beta)}+\frac{1}{(\alpha B)}+\frac{1}{(\alpha\beta)}} \quad . \qquad . \quad (21.33)$$

This quantity is not infinite, as might at first sight appear, if one of the cell frequencies vanishes, because in that case $1-Q^2$ also vanishes; in fact, in such an event $\sigma_Q = 0$.

Standard Error of the Coefficient of Mean Square Contingency.

21.30. The determination of the standard error of the coefficient of mean square contingency is a matter of considerable mathematical complexity, and even when approximations are employed, leads to expressions which are tedious to calculate in practice. For a detailed discussion we must refer the student to the original memoirs (refs. (448) and (489)).

The Rank Correlation Coefficient.

21.31. Unlike most of the parameters we have been considering, the distribution of the rank correlation coefficient is discontinuous, and to that extent resembles the binomial. Very little is known about the distribution except in the important case when the correlation in the universe is zero. The other cases are sometimes treated by assuming a normal continuous distribution in the parent and working from ranks to grades and thence to the product-moment coefficient of correlation by the equations (13.11) and (13.12) of **13.21**; but this procedure is hardly to be recommended.

The case when the correlation in the universe is zero, *i.e.* when all possible permutations of the ranks occur with equal frequency, has to some extent been investigated. It was shown by "Student" in 1907 that the standard deviation of the rank correlation coefficient is given by the simple equation

$$\sigma_\rho = \frac{1}{\sqrt{n-1}} \qquad . \qquad . \qquad . \qquad . \quad (21.34)$$

This cannot be taken to be a standard error in the ordinary way, because the distribution is not normal for small samples. But it has been shown by Hotelling and Pabst (ref. (540)) that for large samples the distribution may be taken to be continuous and normal, whether the universe can be regarded as classified according to a continuous variate or not. The appearance of the normal curve in this connection is peculiar and unexpected, for the distribution in small samples might lead one to expect a bimodal distribution.

21.32. Unfortunately, the rank correlation coefficient is mostly used for samples of 10 to 50, and it is not yet clear whether the latter number is large enough for the normality of the distribution in large samples to be used. It would appear that for samples of 10 or 20, at least, the distribution itself should be obtained, and further research on this subject would be useful.

SUMMARY.

1. The following are the standard errors of the parameters named, the parent universe being assumed normal :—

Variance $\qquad\qquad \sigma^2\sqrt{\dfrac{2}{n}}$

Standard deviation $\qquad\qquad \dfrac{\sigma}{\sqrt{2n}}$

Coefficient of variation $\qquad \dfrac{V}{\sqrt{2n}}\sqrt{1+\dfrac{2V^2}{10^4}}$

Correlation coefficient $\qquad \dfrac{1-r^2}{\sqrt{n}}$

Regression coefficient $\qquad \dfrac{\sigma_1\sqrt{1-r^2}}{\sigma_2\sqrt{n}}$ or $\dfrac{\sigma_{1.2}}{\sigma_2\sqrt{n}}$

2. The standard error of the qth moment measured about the mean is given by

$$\sigma_{\mu_q} = \sqrt{\frac{\mu_{2q} - \mu_q^2 + q^2\mu_2\mu_{q-1}^2 - 2q\mu_{q-1}\mu_{q+1}}{n}}$$

3. The correlation between errors in the qth and rth moments, both measured about the mean, is given by

$$\sigma_{\mu_q}\sigma_{\mu_r}r_{\mu_q\mu_r} = \frac{\mu_{q+r} - \mu_q\mu_r + qr\mu_2\mu_{q-1}\mu_{r-1} - r\mu_{q+1}\mu_{r-1} - q\mu_{q-1}\mu_{r+1}}{n}$$

4. From the results of (2) and (3), and similar results for moments about a fixed point, it is possible to calculate the standard error of any function of the moments.

5. In the normal universe, errors in the mean and standard deviation are uncorrelated.

6. In calculating the standard errors of moments the uncorrected values should be used.

7. It is unsafe to use the formulæ for standard errors appropriate to the normal universe in cases where the universe is suspected to differ from the normal form ; in particular, the formula for the standard error of the standard deviation, $\dfrac{\sigma}{\sqrt{2n}}$, should not be used for parent universes which are markedly lepto- or platy-kurtic.

EXERCISES.

21.1. In the weight distribution of Exercise 6.6, page 111, last column, find the standard error of the standard deviation. Compare it with the value obtained on the assumption that the parent distribution is normal.

21.2. In the same data, compare the ratio of the s.e. of the s.d. to the s.d. with the ratio of the s.e. of the semi-interquartile range to the semi-interquartile range.

21.3. Show that for a normal universe the standard error of the s.d. is less than the standard error of the semi-interquartile range.

21.4. In a sample of 1000 the mean is found to be 17·5 and the standard deviation 2·5. In another sample of 800 the mean is 18 and the standard deviation 2·7. Assuming that the samples are independent, discuss whether the two samples can have come from universes which have the same standard deviation.

21.5. Find the correlation between errors in the mean and standard deviation for the height distribution of 8585 men of Table 6.7, page 94, and do the same for the marriage distribution of Table 6.8, page 96.

21.6. Find the standard errors of the first four seminvariants as calculated from the moments.

21.7. Samples of 10,000 are taken from a normal universe. For what even moments does the standard error of the moment lie within 10 per cent. of the value of that moment?

21.8. For samples of (a) 100, (b) 1000, draw a graph showing how the standard error of the correlation coefficient from a normal universe varies with r.

21.9. (Data quoted by M. F. Hoadley, "Note on the Association of Relative Laterality of Hand and Eye from the Cambridge Anthropometric Data," *Biometrika*, vol. 20B, 1928, p. 401.)

Three experiments were conducted to determine the relationship between laterality of hand and laterality of eye. The correlations between (1) difference of strength of grip and (2) difference in visual acuity were:

$$-0·02410 \quad (3234 \text{ subjects})$$
$$-0·00738 \quad (4003 \text{ subjects})$$
$$+0·02962 \quad (1447 \text{ subjects})$$

Find the standard errors of the three correlation coefficients, and hence show that it cannot be concluded that there is any significant correlation between laterality of hand and laterality of eye.

21.10. Find the standard errors of the partial correlation coefficients of Example 14.1, page 270. Hence state whether any one is not significantly different from zero, and if so, which. For the purpose of this exercise normality may be assumed, although in all probability the actual data do not emanate from a normal universe.

CHAPTER 22.

THE χ^2 DISTRIBUTION.

22.1. In Chapters 19 to 21 we have seen that a knowledge of the sampling distribution of a parameter gives us a means of judging from samples the relationship between fact and theory. For instance, in Example 19.8, page 352, we were able to infer from a knowledge of the binomial distribution that the dice which provided the data were probably biased ; and in Example 20.6, page 386, we could apply a knowledge of the distribution of the mean of samples from a normal population to reject the hypothesis that the mean in the universe was less than 67 inches.

In the present chapter we shall discuss a particular sampling distribution of profound importance in statistical theory, and shall note its applications to the testing of accordance between fact and hypothesis in a wide range of cases.

Cells.

22.2. In what follows we shall consider only data giving the frequencies of individuals falling within various categories. Statistical data, as will have been evident from the examples already given in this book, are very often of this type.

Such data, whether relating to attributes or to continuous variates or to a mixture of both, will in practice be arranged in compartments. For example, in the association table on page 40 there are four compartments, corresponding to the four ultimate classes. In the table of frequencies within various height ranges (Table 6.7, p. 94), each range determines a compartment, and the data consist of 8585 individuals distributed in 21 groups.

It is convenient to have a name for these compartments. We shall call them **cells**. The frequency falling in a cell will be referred to as the **cell frequency**.

One and the same table may contain frequencies of more than one order, and frequencies of different orders must be kept distinct. Thus an association table has four cells with frequencies of the second order and two sets of two (the border frequencies) of the first order. A $p \times q$ contingency table has pq cells of the second order (to condense our terminology) and a set of p and a set of q of the first order. Each such set must be considered by itself. The tests of this chapter are applicable to any homogeneous set, but not to a " mixed " set comprising cells of different orders.

22.3. We shall denote the number of cells in the presentation of a set of data by n, and the cell frequency occurring in the rth cell by $\tilde{m}_r$. Thus, in the table of page 94 we have, numbering the cells downwards:

413

$$\tilde{m}_1 = 2$$
$$\tilde{m}_2 = 4$$
$$\tilde{m}_3 = 14$$
. . . .
$$\tilde{m}_{21} = 2$$

22.4. In the class of cases we shall consider, we wish to compare the actual values $\tilde{m}$ with the cell frequencies which would exist if a particular hypothesis H were exactly verified. These latter values we shall denote by the letter m, so that the theoretical frequency in the rth cell is m_r.

The cell frequencies m_r are sometimes referred to as the " expected " values on the hypothesis H. This is rather a special use of the word " expected," in the sense we have already given, namely, that the m_r's assume the values which they would take if the hypothesis were exactly verified for the particular set of data.

We shall write :

$$x_r = \tilde{m}_r - m_r \qquad . \qquad . \qquad . \qquad (22.1)$$

so that the x_r's are the excesses of the actual over the expected frequencies.

Clearly the quantities x embody all the information in the data about the discrepancies between theory and fact. If the x's are all zero, fact and theory are in perfect agreement. If the x's are large, the agreement is poor.

Example 22.1.—As a simple example let us consider the 2×2 contingency table of Example 3.5, page 40. Numbering the cells from left to right we have :

$$\tilde{m}_1 = 276, \qquad \tilde{m}_2 = 3$$
$$\tilde{m}_3 = 473, \qquad \tilde{m}_4 = 66$$

Now let our hypothesis H be that inoculation and exemption from attack are independent. If this be so, the expected frequencies are :

$$m_1 = 255 \cdot 5, \qquad m_2 = 23 \cdot 5$$
$$m_3 = 493 \cdot 5, \qquad m_4 = 45 \cdot 5$$

and hence we have :

$$x_1 = \tilde{m}_1 - m_1 = 20 \cdot 5, \qquad x_2 = -20 \cdot 5$$
$$x_3 = -20 \cdot 5, \qquad x_4 = 20 \cdot 5$$

The x's are, in fact, in this particular case, the numbers we referred to in Chapter 3 as δ-numbers. We have already considered them as reflecting the divergence of fact from theory.

Constraints.

22.5. In the example we have just considered, one important effect is to be noted, viz. that when we have calculated one independent frequency, say m_1, the other three follow arithmetically from the fact that the two frequencies in any row or column must add up to the border frequency in that row or column.

In fact, we have :

$$\left. \begin{array}{l} x_1 + x_2 = 0 \\ x_1 + x_3 = 0 \\ x_2 + x_4 = 0 \end{array} \right\} \qquad . \qquad . \qquad . \qquad . \qquad (22.2)$$

We need not add $x_3 + x_4 = 0$, since this is given by the last two equations in conjunction with the first. There are only three independent equations.

Thus, whatever our hypothesis H may be, the conditions of the problem impose limitations, expressed by the equations (22.2), on the way in which the m's and the x's may be chosen. If one m or one x is fixed by H, the other three are determinate in accordance with the conditions of the data themselves.

Similarly, suppose we wished to examine the height data of page 94 in the light of the hypothesis that the parent distribution, of which this is a sample, is normal with given mean and standard deviation. With the aid of the table of the probability integral we can determine the cell frequencies on this hypothesis ; but again the problem imposes a limitation on the way in which the theoretical cell frequencies are assigned, namely, that they must add up to the total number 8585 of the sample. When 20 frequencies are fixed, the other is determined by mere arithmetic.

22.6. In general, when the conditions of the problem impose limitations of this kind on the number of cell frequencies which may be fixed by H we say, borrowing an expression from Statics, that they impose **constraints**. In the example of the 2×2 contingency table there were three independent constraints, expressed by the equations (22.2). In the case of the height distribution there is one constraint expressed by the fact that the sum of the cell frequencies must be 8585.

Linear Constraints.

22.7. Constraints which involve linear equations in the cell frequencies (*i.e.* equations containing no squares or higher powers of the frequencies) are called **linear constraints**. The two instances above are of this type. Linear constraints are of paramount importance, and we shall shortly confine our attention to them alone.

Degrees of Freedom.

22.8. We denote the number of independent constraints in a set of data by κ. We then define the number ν by the simple equation

$$\nu = n - \kappa$$

and call ν the number of **degrees of freedom** of the aggregate of cells. It is the number of cell frequencies which can be assigned at will, the remaining κ following from the conditions to which the data are subject.

Thus, for the 2×2 table $\kappa = 3$ and $\nu = 1$, for, as we have seen, the fixing of one cell frequency fixes them all. For the height distribution $\kappa = 1$, $\nu = 20$.

Example 22.2.—Let us find the number of degrees of freedom of a $p \times q$ contingency table.

The constraints of such a table are similar to those of the 2×2 table. Thus the sum of the cell frequencies in each row is determined as being the border frequency in that row, and similarly for the columns. Hence each of the p columns and q rows imposes a constraint. From the total $p + q$ constraints we must, however, subtract one, for they are not algebraically independent ; there is one relation between them, expressed by the fact that the sum of the border column equals the sum of the border row, namely, the total frequency N.

Hence there are $p + q - 1$ independent linear constraints.   Hence,

$$\nu = n - \kappa$$
$$= pq - (p + q - 1)$$
$$= (p - 1)(q - 1)$$

We might have got this result more directly by considering that the cell frequencies in the first $p - 1$ columns and $q - 1$ rows are determinable at will, the rest following automatically from the border frequencies. Hence the number of degrees of freedom, being the number of cells which can be so filled, is $(p - 1)(q - 1)$ as before.

22.9.   Now let us consider a set of data arranged in n cells, the total frequency being N.

The theoretical frequency in the rth cell is m_r.   This means that the chance of an individual falling into this cell is $\frac{m_r}{N}$, and the chance of its not doing so is $\left(1 - \frac{m_r}{N}\right)$.   We may regard the actual frequencies $\tilde{m}$ as having been arrived at by distributing the N individuals among the n cells in such a way that the chance of an individual falling into the rth cell is $\frac{m_r}{N}$.   Hence the probability that of the N individuals, $\tilde{m}_r$ fall into the rth cell and the remainder elsewhere is the term

$$\left(\frac{m_r}{N}\right)^{\tilde{m}_r}\left(1 - \frac{m_r}{N}\right)^{N - \tilde{m}_r}$$

in the binomial

$$\left\{\frac{m_r}{N} + \left(1 - \frac{m_r}{N}\right)\right\}^N$$

Thus, this binomial will give us the relative frequencies of the various values which $\tilde{m}_r$ can take in different samples, of which the actual data form one.

If N is fairly large and $\frac{m_r}{N}$ is not small, this distribution is approximately normal with mean m_r.   That is to say, $\tilde{m}_r$ is distributed normally about a mean m_r, or x_r is distributed normally about zero mean.

Definition of χ^2.

22.10.   We now define the quantity χ^2 by the equation

$$\chi^2 = S\left(\frac{x_r^2}{m_r}\right) = S\left\{\frac{(\tilde{m}_r - m_r)^2}{m_r}\right\}   \qquad . \qquad . \qquad (22.3)$$

the summation being taken over the n cells.

The student can verify for himself that this definition is consistent with that given in equation (5.4), page 68, for the particular case of divergence from independence in a contingency table.

We can write χ^2 in a slightly different form. For

$$\chi^2 = S\left\{\frac{(\tilde{m}_r - m_r)^2}{m_r}\right\} = S\left(\frac{\tilde{m}_r{}^2}{m_r}\right) - 2S\left(\frac{\tilde{m}_r m_r}{m_r}\right) + S\left(\frac{m_r{}^2}{m_r}\right)$$

$$= S\left(\frac{\tilde{m}_r{}^2}{m_r}\right) - 2S(\tilde{m}_r) + S(m_r)$$

$$= S\left(\frac{\tilde{m}_r{}^2}{m_r}\right) - N\ . \qquad . \qquad . \qquad . \quad (22.4)$$

This corresponds to equation (5.7), page 69.

22.11. If $\chi^2 = 0$ all the x's are zero, and hence the actual cell frequencies coincide with the expected cell frequencies. On the other hand, if some or all of the x's are large, χ^2 will be large.

It will thus be evident that χ^2 affords a measure of the correspondence between fact and theory. It must not be forgotten, however, that it ignores the signs of the x's and hence takes no cognisance of certain information which those signs may convey. We shall take up this point again later.

22.12. If the use of χ^2 is to be satisfactory, we must be able to distinguish significant values from those which may have arisen by sampling fluctuations. This leads us to inquire what is the probability of getting a particular value of χ^2 from a set of $\tilde{m}_r$'s chosen at random, and this in turn leads to the question : What is the sampling distribution of χ^2 ?

We shall not give a proof here of the important answer to this question, but shall content ourselves with quoting it and indicating briefly the method by which it is obtained.

We have already seen that the sum of n normally distributed variates is itself normally distributed (12.8). The sum of the *squares* of n normal variates is not so distributed, however. In fact, the sum of the squares of n normal variates, drawn from a universe with unit standard deviation, is distributed in a form given by the equation

$$y = y_0 e^{-\frac{\Sigma^2}{2}} \Sigma^{n-1} \qquad . \qquad . \qquad . \qquad . \quad (22.5)$$

where Σ^2 is the sum in question.

Now it has already been shown that under the conditions assumed the x's are each distributed normally about zero mean, and it may be shown further that χ^2 may be regarded as the sum of the squares of ν variates each distributed normally with unit s.d. and about a zero mean. Hence the distribution of χ^2 is given by

$$y = y_0 e^{-\frac{\chi^2}{2}} \chi^{\nu-1} \qquad . \qquad . \qquad . \qquad . \quad (22.6)$$

22.13. It follows, as in **20.8**, that if we take a random set of $\tilde{m}$'s and calculate χ^2 from them, the probability of getting a value of χ^2 as great as, or greater than, this observed value $\chi_0{}^2$, is the area of the curve (22.6) to the right of the ordinate at χ_0 divided by the total area of the curve ; or, in the language of the integral calculus,

27

$$P = \dfrac{\displaystyle\int_{\chi_0}^{\infty} y_0 e^{-\frac{\chi^2}{2}} \chi^{\nu-1} d\chi}{\displaystyle\int_0^{\infty} y_0 e^{-\frac{\chi^2}{2}} \chi^{\nu-1} d\chi} \qquad . \qquad . \qquad . \qquad (22.7)\ [1]$$

The curve, as we shall see later, extends from 0 to $+\infty$, which accounts for the limits of the integral in the denominator of the above expression.

Tabulation of P for the χ^2 Distribution.

22.14. The rather formidable result of equation (22.7) need occasion no alarm to the student who is unacquainted with the notation and methods of the integral calculus. The function P has been tabulated for certain ranges of ν and χ^2 in the same way as the probability for the normal curve, and the tables are in most cases sufficient for the practical application of the results of the present chapter.

Tables for $\nu = 1$ are given at the end of this book (Appendix Tables 4A and 4B). Tables for $\nu = 2$ to $\nu = 29$ are given in "*Tables for Statisticians and Biometricians, Part I*," and in the same book are supplementary tables for ranges outside those limits.[2]

For most practical purposes it is not necessary to calculate P to any great degree of accuracy, and the diagram in the Appendix has been drawn to obviate the use of the tables. In this diagram (fig. A1) curves have been drawn to show the relationship between ν and χ^2 for various values of P. The use of the diagram will be apparent from the examples below.

22.15. It is desirable to point out that other writers have used different letters to denote the number of degrees of freedom. Karl Pearson, in the tables to which we have just referred, used the number n', which is one more than our ν. R. A. Fisher writes n instead of our ν, so that we have :

$$\nu = n' - 1 \ (\text{Pearson}) = n \ (\text{Fisher})$$

We have thought it desirable to introduce the symbol ν in order to avoid confusion with the use of n' and n as numbers in a sample or in a universe.

The χ^2 Test of Significance when the Theoretical Cell Frequencies are known *a priori*.

22.16. Armed with the tables of P, or the diagram of the Appendix, we can now proceed as follows :—

---

[1] The actual values of P are, expanding this integral,

$$P = \sqrt{\frac{2}{\pi}} \int_{\chi}^{\infty} e^{-\frac{1}{2}\chi^2} d\chi + \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}\chi^2} \left(\frac{\chi}{1} + \frac{\chi^3}{1.3} + \frac{\chi^5}{1.3.5} + \ldots + \frac{\chi^{\nu-2}}{1.3.5\ldots(\nu-2)} \right)$$

if ν is odd

$$= e^{-\frac{1}{2}\chi^2} \left(1 + \frac{\chi^2}{2} + \frac{\chi^4}{2.4} + \frac{\chi^6}{2.4.6} + \ldots + \frac{\chi^{\nu-2}}{2.4.6\ldots(\nu-2)} \right)$$

if ν is even

The first term of the first series may be obtained from the probability integral.

[2] The work in the introduction to these Tables is inaccurate in some cases, particularly in the treatment of contingency tables, owing to the use of the wrong number of degrees of freedom.

Having decided on the hypothesis to be tested, we calculate from it the theoretical frequencies m_r. (For the present we assume that this can be done without reference to the observed frequencies $\tilde{m}_r$. The contrary case will be considered later.)

From the m_r's and the $\tilde{m}_r$'s we calculate χ^2 according to (22.3) or (22.4). We also ascertain ν.

Then, from the tables, we find the value of P corresponding to these values of χ^2 and ν.

The value P gives us the probability that on random sampling we should get a value of χ^2 as great as, or greater than, the value actually obtained.

Now, if P is small, our data give us an improbable value of χ^2. Thus we have the alternative conclusions that either (a) an improbable event has occurred, or (b) that the divergence of fact from theory is significant of some real effect and cannot be attributed to fluctuations of sampling. The smaller P is, the more we incline to the latter alternative ; if we do decide to adopt it, the inferences we draw will depend on the nature of the problem. Sometimes it will lead us to reject our hypothesis. Sometimes it will lead us to suspect our sampling technique.

The following examples will illustrate the type of reasoning involved in applying the χ^2 test.

Example 22.3.—In some experiments on dice-throwing W. F. R. Weldon rolled 12 dice 26,306 times, observing at each throw the number of dice recording a 5 or a 6.

If the dice are unbiased, the chance of getting a 5 or a 6 with one die is $\frac{1}{3}$. Hence the chances with 12 dice of getting 12 5's or 6's, 11 5's or 6's, etc., are the successive terms in the binomial $\left(\frac{1}{3} + \frac{2}{3}\right)^{12}$. Hence the theoretical frequencies in 26,306 throws are the terms in $26,306 \left(\frac{1}{3} + \frac{2}{3}\right)^{12}$. These are our m_r's.

The following table shows the actual $(\tilde{m}_r)$ and the theoretical (m_r) frequencies, together with the values of $\dfrac{(\tilde{m}_r - m_r)^2}{m_r}$:—

TABLE 22.1.—*12 Dice thrown 26,306 Times, a Throw of 5 or 6 reckoned a Success.*

Number of Successes.	Observed Frequency $(\tilde{m})$.	Theoretical Frequency (m).	$\tilde{m} - m$ (x).	$\dfrac{(\tilde{m} - m)^2}{m}$.
0	185	203	− 18	1·596
1	1,149	1,217	− 68	3·800
2	3,265	3,345	− 80	1·913
3	5,475	5,576	− 101	1·829
4	6,114	6,273	− 159	4·030
5	5,194	5,018	+ 176	6·173
6	3,067	2,927	+ 140	6·696
7	1,331	1,254	+ 77	4·728
8	403	392	+ 11	0·309
9	105	87	+ 18	3·724
10 and over	18	14	+ 4	1·143
Totals	26,306	26,306	0	35·941

Hence $\chi^2 = 35\cdot941$, and $\nu =$ one less than the number of cells $= 10$.

From the "*Tables for Statisticians and Biometricians*" we have, when $\nu = 10$ ($n' = 11$),

$$P = 0 \cdot 000857 \quad \text{for} \quad \chi^2 = 30$$
$$P = 0 \cdot 000017 \quad \text{for} \quad \chi^2 = 40$$

Evidently when $\chi^2 = 35 \cdot 941$, P will be extremely small. If we want to evaluate it exactly we can proceed by the methods given in the Tables. In fact $P = 0 \cdot 000086$.

Alternatively, from the diagram we see that when $\chi^2 = 35 \cdot 94$ and $\nu = 10$, the value of P lies slightly below $0 \cdot 0001$, for the point with ordinate 10 and abscissa $35 \cdot 94$ lies close to, but below, the curve labelled $P = 0 \cdot 0001$.

Thus the probability that, on random sampling, we should get an equally or less close approach to the observed value of χ^2 is less than one in 10,000.

We may therefore say that the correspondence between theory and fact is very poor. The extreme improbability of the observed event enables us to say with some confidence that the divergence between the two is significant, and hence that either our sampling technique or our hypothesis is at fault. Now in this experiment Weldon took particular care with the dice-throwing, and we may regard it as unlikely that there was anything seriously wrong with the randomness of the sampling. We are therefore led to doubt our hypothesis that the dice were unbiased.

Briefly, then, the χ^2 test suggests that the dice were biased.

Example 22.4.—(Data from ref. (74).) The following table shows the result of inoculation against cholera on a certain tea estate :—

TABLE 22.2.

	Not-attacked.	Attacked.	Total.
Inoculated	431 (427·7)	5 (8·3)	436
Not-inoculated.	291 (294·3)	9 (5·7)	300
Total	722	14	736

We shall explain the figures in brackets presently. The question on which we want to throw light is : Is there any significant association between inoculation and attack ?

To answer this, let us take for our hypothesis H the supposition that they are independent. If this is so, the expected frequencies, calculated in the manner of Chapter 3, are those given in brackets. These we take to be the m_r's, the $\tilde{m}_r$'s being the actual frequencies. We then have :

$$\chi^2 = (3 \cdot 3)^2 \left\{ \frac{1}{427 \cdot 7} + \frac{1}{8 \cdot 3} + \frac{1}{294 \cdot 3} + \frac{1}{5 \cdot 7} \right\} = 3 \cdot 27$$

and

$$\nu = 1$$

From Appendix Table 4B, $P = 0 \cdot 0706$.

Thus if H is true, our data give a result which would be obtained about seven times in a hundred trials. This is infrequent, but not very infrequent. Moreover, the theoretical frequencies in the "attacked" column are not very large. We should therefore be unjustified in rejecting H on this evidence, but we can say that the data lend some colour to the supposition that H is not correct.

To sum up, the χ^2 test shows that the data incline us, though not strongly, to the belief that inoculation and attack are associated.

Example 22.5.— (Imaginary data.) An investigator into chocolate consumption divided the United Kingdom into eight areas and took a random sample from each, the individuals so obtained being classified as consumers or non-consumers of chocolate. His results were as follows:—

<center>TABLE 22.3.</center>

Area Number .	1.	2.	3.	4.	5.	6.	7.	8.	Total.
Consumers . .	56 (55)	87 (81)	142 (152)	71 (69)	88 (90)	72 (72)	100 (95)	142 (144)	758
Non-consumers .	17 (18)	20 (26)	58 (48)	20 (22)	31 (29)	23 (23)	25 (30)	48 (46)	242
Total .	73	107	200	91	119	95	125	190	1000

Do these results suggest that the consumption of chocolate varies from place to place ?

Let us take as our hypothesis H the supposition that it does not, *i.e.* that the two attributes in the above table are independent. The theoretical frequencies m_r are then those shown in brackets, and we have:

$$\chi^2 = \frac{1^2}{55} + \frac{6^2}{81} + 14 \text{ similar terms}$$

$$= 6\cdot 28$$

The table has two rows and eight columns, and hence $\nu = (2-1)(8-1) = 7$. From the diagram of the Appendix, the point whose abscissa is 6·28 and ordinate 7 lies between the lines $P = 0\cdot 75$ and $P = 0\cdot 5$, very near the latter; or alternatively, from the "*Tables for Statisticians and Biometricians*" for $\nu = 7$ $(n' = 8)$,

$$\text{if } \chi^2 = 6, \qquad P = 0\cdot 539750$$
$$\text{if } \chi^2 = 7, \qquad P = 0\cdot 428880$$

Hence, for $\chi^2 = 6\cdot 28$, $P = 0\cdot 51$ approximately.

Thus there is no cause to suspect our hypothesis, and the data do not suggest that the consumption of chocolate varies from place to place, at least so far as this test is concerned.

Properties of the χ^2 Distribution.

22.17. The curves

$$y = y_0 e^{-\frac{\chi^2}{2}} \chi^{\nu-1}$$

and the probability function P derived from them, have several interesting properties which are worth noticing. As χ^2 is essentially positive, we consider only positive values of the variate.

(a) In the first place, it will be seen that when $\nu = 1$ the curve is the normal curve with unit standard deviation, for positive values of the variate. Thus the test for $\nu = 1$ may be reduced to testing the significance of deviations of a normally distributed variate.

(b) When $\nu > 1$ the curve is of the single-humped type. It is tangential to the x-axis at the origin ($\chi^2 = 0$), rises to a maximum where $\chi^2 = \nu - 1$ and then falls more slowly to zero as χ^2 increases indefinitely. It is thus skew to the right.

(c) As ν increases, the curve becomes more and more symmetrical. In fact, when ν is large, $\sqrt{2\chi^2}$ is distributed approximately normally about a mean $\sqrt{2\nu - 1}$ with unit standard deviation. This result, due to R. A. Fisher, enables us to dispense with tables of P for large values of ν, say $\nu > 30$, and to use the probability integral instead. In practice large values of ν are rather infrequent.

Example 22.6.—To find P when $\chi^2 = 64$ and $\nu = 41$.

We know that $\sqrt{2\chi^2}$ is distributed normally about mean $\sqrt{82 - 1} = 9$ with unit standard deviation. When $\chi^2 = 64$, $\sqrt{2\chi^2} = 11\cdot314$, which therefore has a deviation $2\cdot314$ to the right of the mean. Hence we have to find the area of the probability curve to the right of the ordinate which is $2\cdot314$ units to the right of the mean. From Appendix Table 2 this is seen to be $0\cdot0104$ approximately.

Conditions for the Application of the χ^2 Test.

22.18. We may conveniently bring together at this point the various precautions which should be observed in applying the χ^2 distribution to a test of significance.

(a) In the first place, N must be reasonably large. Otherwise the x's are not normally distributed.

This is a condition which is almost always fulfilled in practice. It is difficult to say exactly what constitutes largeness, but as an arbitrary figure we may say that N should be at least 50, however few the number of cells.

(b) No theoretical cell frequency should be small. Here again it is hard to say what constitutes smallness, but 5 should be regarded as the very minimum, and 10 is better.

In practice, data not infrequently contain cell frequencies below these limits. As a rule the difficulty may be met by amalgamating such cells into a single cell. Thus, in Example 22.3 above, the theoretical numbers of throws with 10, 11 and 12 successes are (to the nearest integer) 13, 1 and 0. Instead of putting each into a separate cell we have run them together into one cell " 10 and over."

(c) The constraints must be linear. The reason for this condition has not emerged explicitly in the foregoing because we omitted the stage in the proof of the χ^2 distribution at which it occurs.

22.19. To these three conditions we may add the following remarks, which should also be borne in mind when the χ^2 test is being used.

(a) The χ^2 test tells us the probability of getting, on a random sample, a value of χ^2 equal to or higher than the actual value. If this probability is small we are justified in suspecting a significant divergence between theory and experiment.

We cannot proceed, however, in the reverse direction and say that if P is not small our hypothesis is proved correct. All that we can say is that the test reveals no grounds for supposing the hypothesis incorrect; or alternatively, that so far as the χ^2 test is concerned, data and hypothesis are in agreement.

(b) Nor do only small values of P lead us to suspect our hypothesis or our sampling technique. A value of P very near to unity may also do so.

This rather surprising result arises in this way: a large value of P normally corresponds to a small value of χ^2, that is to say a very close agreement between theory and fact. Now such agreements are rare—almost as rare as great divergences.

We are just as unlikely to get very good correspondence between fact and theory as we are to get very bad correspondence and, for precisely the same reasons, we must suspect our sampling technique if we do. In short, very close correspondence is *too good to be true*.

The student who feels some hesitation about this statement may like to reassure himself with the following example. An investigator says that he threw a die 600 times and got exactly 100 of each number from 1 to 6. This is the theoretical expectation, $\chi^2 = 0$ and $P = 1$, but should we believe him? We might, if we knew him very well, but we should probably regard him as somewhat lucky, which is only another way of saying that he has brought off a very improbable event.

22.20. At this point we can resume a topic which we laid on one side in **22.11**, namely the signs of the x's, which are ignored by χ^2.

It may happen that χ^2 has quite a moderate value and P is not small when all the positive x's are on one side of the mode of the theoretical distribution and all the negative x's on the other. There will thus be a consistent " shift " of the $\bar{m}$'s one way or the other from the m's. This may give us a value of the mean quite outside the limits of sampling. Again, if the x's are all negative in the cells farthest removed from the mean, the standard deviation may show an almost impossible divergence from expectation.

Thus, although the χ^2 test may reveal no cause to suspect the hypothesis, a closer examination of the x's may.

Example 22.7.—Consider the following dice data (Table 22.4) (Weldon, see p. 351).

Now, in this example, all the x's are negative up to 5 successes, positive from 6 to 10 successes, and negative again for 11 to 12 successes. This is almost one of the cases we referred to earlier in this section.

We have, in fact, already found (Example 19.3, page 352) that the mean deviates from the expected value by 5·13 times the standard error.

TABLE 22.4.—*12 Dice thrown 4096 times, a Throw of 4, 5 or 6 Points reckoned a Success.*

Number of Successes.	Observed Frequency $(\bar{m})$.	Expected Frequency (m). $4096(\frac{1}{2}+\frac{1}{2})^{12}$	$\bar{m}-m$ (x).	$\dfrac{(\bar{m}-m)^2}{m}$
0	0	1	-1	1·0000
1	7	12	-5	2·0833
2	60	66	-6	0·5455
3	198	220	-22	2·2000
4	430	495	-65	8·5354
5	731	792	-61	4·6982
6	948	924	24	0·6234
7	847	792	55	3·8194
8	536	495	41	3·3960
9	257	220	37	6·2227
10	71	66	5	0·3788
11	11 } 11	12 } 13	-1 } -2	0·3077
12	0	1	-1	
Totals	4096	4096	0	33·8104 $=\chi^2$

From the tables we find:

ν	n'	χ^2	P
12	13	30	0·002792
12	13	40	0·000072

Hence, by simple interpolation for $\chi^2 = 33.8104$, $P = 0.0018$.

As a matter of fact, simple interpolation is of very little value for small values of P (*cf.* 24.12), and this value is wide of the mark, the true value being 0·00072. A better idea is to be gained from the Appendix diagram, from which it is seen that P lies between 0·001 and 0·0001. In any case, the value of P is small, but not overwhelmingly small.

From the extended tables of the normal integral in "*Tables for Statisticians and Biometricians, Part I,*" we have:

Greater fraction of the area of a normal
    curve for a deviation 5·13 .   .   .   0·9999998551
Area in the tail of the curve .   .   .   0·0000001449
Area in both tails .   .   .   .   0·0000002898

so that the probability of getting such a deviation (+ or −) on random sampling is only about 3 in 10,000,000.

Comparing this with the value of P, we see that the data are really more divergent from theory than the χ^2 test would lead us to suppose.

22.21. Hence, if the signs of the x's show any marked peculiarities, it is as well to apply as many supplementary tests as are available, and not to rely on the χ^2 test alone. Such tests would include those for the significance of the mean and standard deviation, which we have already discussed.

Levels of Significance.

22.22. In the examples we have given above, our judgment whether P was small enough to justify us in suspecting a significant difference between

fact and theory has been more or less intuitive. Most people would agree, in Example 22.3, that a probability of only 0·0001 is so small that the evidence is very much in favour of the supposition that the dice were biased. But we shall not always get such a decisive result. Suppose we had obtained $P = 0·1$, so that the odds against the event are nine to one: Is this value small enough to lead us to suspect the dice? If it is not, would $P = 0·01$ be small enough? Where, if anywhere, can we draw the line?

The odds against the observed event which influence a decision one way or the other depend to some extent on the caution of the investigator. Some people (not necessarily statisticians) would regard odds of ten to one as sufficient. Others would be more conservative and reserve judgment until the odds were much greater. It is a matter of personal taste.

22.23. There are, however, two values of P which are widely used to provide a rough line of demarcation between acceptance and rejection of the significance of observed deviations. These values are $P = 0·05$ and $P = 0·01$, and are said to define 5 per cent. and 1 per cent. *levels of significance.* The value $P = 0·001$, *i.e.* the 0·1 per cent. level, is also used. If we choose to adopt these levels, our attention will be focused, not as heretofore on the actual value of P, but on the fact whether it falls above or below the levels of significance. To facilitate the investigation of this aspect of the matter, R. A. Fisher has prepared tables (published in his " *Statistical Methods for Research Workers* ") in a different form from those of " *Tables for Statisticians and Biometricians,*" which are due to W. Palin Elderton. The latter, as we have mentioned, give the values of P corresponding to given values of χ^2 and ν. Fisher's tables give χ^2 corresponding to given values of ν and P, and among those values are $P = 0·05$ and $P = 0·01$—the significance levels.

The diagram of the Appendix expresses a similar point of view, and gives the curves of relationship between χ^2 and ν for constant values of P, or, in short, the contour lines of the surface

$$P = F(\chi^2, \nu)$$

The diagram gives the 5 per cent. and 1 per cent. lines and also those corresponding to the smaller probabilities $P = 0·001$ and $0·0001$, *i.e.* the 0·1 per cent. and the 0·01 per cent. levels.

A value of P less than 0·05 will be said to fall *below* the 5 per cent. level of significance, and so on.

Example 22.8.—Let us consider the data of Exercise 3.11. In experiments on the Spahlinger anti-tuberculosis vaccine the following results were obtained. (As before, the figures in brackets are the independence values.)

	Died or Seriously Affected.	Unaffected or Not Seriously Affected.	Total.
Inoculated	6 (8·87)	13 (10·13)	19
Not inoculated or inoculated with control media	8 (5·13)	3 (5·87)	11
Total	14	16	30

Here, $$\chi^2 = 4\cdot75 \quad \text{and} \quad \nu = 1$$

From Appendix Table 4B we have $P = 0\cdot029$ approximately.

Alternatively, from Fisher's table we have, when $\nu = 1$,

for $P = 0\cdot05 \qquad \chi^2 = 3\cdot841$

and

for $P = 0\cdot01 \qquad \chi^2 = 6\cdot635$

so that, from either table, P lies between the 5 per cent. level of significance and the 1 per cent. level.

If, therefore, we take the 5 per cent. level as appropriate to this case, the results are significant ; but if we are more conservative and take the 1 per cent. level, the results are not significant. In this particular case the position is complicated by the relative smallness of the theoretical cell frequencies.

The Additive Property of χ^2.

22.24. It sometimes happens, by the repetition of experiments or otherwise, that we have a number of tables for similar data from different fields. The values of P for each may not be entirely conclusive. The question then arises whether we cannot obtain a value of P for the aggregate, telling us what is the probability of getting, by random sampling, a series of divergences from theory as great as or greater than those observed.

The question is usually answered by pooling the results to form a single table. But, apart from the fact that this is not always possible, we have already seen (Chapter 4) that pooling is likely to introduce fallacies. A better method is to proceed in accordance with the following general rule.

22.25. Suppose we have a number of groups of data, each furnishing a χ^2 and a ν. Add together all the χ^2's to form a single value χ_1^2, and all the ν's to form a single value ν_1. The χ^2 test may then be applied to χ_1^2 and ν_1 as if they came from a single set of cells.

The validity of this rule will be evident when we consider how the χ^2 test was arrived at. The variate x in every cell is normally distributed about a mean m, and χ_1^2 is the sum of the squares of quantities like $\dfrac{x^2}{m}$ just as χ^2 was. This, together with the linearity of the constraints, which remains, was the essential part of the proof of the χ^2 distribution, and hence the test remains true for χ_1^2 and ν_1.

Example 22.9.—In Example 22.4 (inoculation against cholera on a certain tea estate) we saw that the χ^2 test, although suggesting that inoculation had some effect in immunising, did not allow us to place any great confidence in such a conclusion. The following data give χ^2 and P for six estates, including the one we have already discussed :—

χ^2.	P.
9·34	0·0022
6·08	0·014
2·51	0·11
3·27	0·071
5·61	0·018
1·59	0·21
Total 28·40	

Here only one value of P is less than 0·01, and we might be inclined to doubt whether the association between inoculation and immunity is real. Let us, however, add the values of χ^2 and of ν. We get $\chi_1^2 = 28\cdot40$ and $\nu_1 = 6$, there being one degree of freedom from each of the six tables.

From the diagram of the Appendix we see that for these values P is slightly below the value 0·0001. If we require greater accuracy, from the tables we have :

χ^2.	P.
28	0·000094
29	0·000061

Whence by interpolation $P = 0\cdot00008$ approximately, *i.e.* we should expect to get a χ^2 as great as this only 80 times in a million. We can, therefore, regard the results, taken together, as significant with a high degree of confidence.

Estimation of Theoretical Frequencies from the Data.

22.26. Our theoretical frequencies m may be calculated partly on the basis of information from the data, partly on *a priori* grounds. Thus, in the dice-throwing data of Example 22.3, our hypothesis that the dice were unbiased enabled us to say that the chance of getting a 5 or a 6 was $\frac{1}{3}$, and hence that the chances with 12 dice were the terms in $26{,}306\,(\frac{2}{3} + \frac{1}{3})^{12}$. Here we take only the value of N, the total frequency, from the data.

In the association and contingency tables, the values of row and column totals, as well as N, are taken from the data and we assume *a priori* that the attributes are independent.

It may be, however, that we draw further information from the data themselves in fixing the theoretical frequencies. In such cases an important modification is necessary in the previous methods of work, for the number of degrees of freedom is further restricted by each piece of information drawn from the data, as we have already seen for contingency tables.

22.27. Consider, for example, the dice-throwing data of Example 22.3. We have already seen that the dice were probably biased, so that the chance of a success was not $\frac{1}{3}$. What, then, was it ?

To answer this question we can only appeal to the data. The proportion of 5's and 6's in the total number of throws of individual dice $(26{,}306 \times 12)$ was 0·3377. Let us therefore take this to be an estimate of the true probability. We can be confident that it will be somewhere very close, owing to the large number in the sample. The theoretical frequencies will then be the terms in $26{,}306\,(0\cdot6623 + 0\cdot3377)^{12}$.

To take a second case : consider the height distribution of Table 6.7, page 94. We have already had reason to suspect that this is a sample from a normal population. If we suppose this hypothesis to be correct, the question arises, What is the mean and standard deviation of the universe ? Here again we must estimate these quantities from the data, in the manner of Chapter 20.

22.28. We shall denote values of the theoretical frequencies which are calculated from parameters estimated from the data by the letter m', and the value of χ^2 calculated from them by χ'^2, so that we have :

$$\chi'^2 = S\left\{\frac{(\tilde{m} - m')^2}{m'}\right\}$$

Now, χ'^2 is an estimate of χ^2 and, if the m''s are close to the m's, χ'^2 will be close to χ^2. χ'^2 is made up of two parts, one measuring the divergence between theory and fact, the other due to errors of estimation of χ^2.   If the second is small compared with the first, we may expect that the χ^2 test, applied with χ'^2 instead of the unknown χ^2, will continue to reveal significant differences between theory and fact where such exist.

22.29.   The question as to the precise conditions under which the test is applicable for such cases has not been completely answered, but it has been shown that, if the cell frequencies are large, the test still applies subject to the following conditions :—

(a) The number of degrees of freedom must be reduced by unity for each constant of the universe which is estimated from the data.

(b) The estimates must be of the type known as " efficient."

We shall not be able in this Introduction to go into the theory of this important class of estimate, but it will be sufficient if we indicate that the estimates of the mean of a normal universe, and the parameter m of the Poisson distribution, are " efficient " if calculated in the ordinary way, i.e. by taking the value of the parameter in the sample to be the value of the parameter in the universe.

Example 22.10.—Reverting to the data of Example 22.3, let us estimate the true chance of getting a 5 or a 6 from the data themselves. The frequency of the successful event is 0·3377 of the whole.   This is an " efficient " estimate of the chance.   The following table gives the observed frequencies and the theoretical frequencies calculated from the formula $26{,}306 \ (0{\cdot}6623 + 0{\cdot}3377)^{12}$:—

TABLE. 22.5.—*12 Dice thrown 26,306 Times, a Throw of 5 or 6 reckoned a Success.*

Number of Successes.	Observed Frequency ($\tilde{m}$).	Theoretical Frequency (m').	$\tilde{m} - m'$.	$\dfrac{(\tilde{m} - m')^2}{m'}$
0	185	187	− 2	0·021
1	1,149	1,146	3	0·008
2	3,265	3,215	50	0·778
3	5,475	5,465	10	0·018
4	6,114	6,269	− 155	3·832
5	5,194	5,115	79	1·220
6	3,067	3,043	24	0·189
7	1,331	1,330	1	0·001
8	403	424	− 21	1·040
9	105	96	9	0·844
10 and over	18	16	2	0·250
Total	26,306	26,306	0	8·201

Thus $\chi^2 = 8{\cdot}201$.   There are 11 cells, with one linear constraint.   We have also fitted one constant from the data, and hence we must take $\nu = 9$.

From the diagram of the Appendix we then see that P is very close to 0·50.

From the tables, for $\nu = 9$ or $n' = 10$, we have:

χ^2.	P.
8	0·5341
9	0·4373

so that $P = 0·51$ approximately.

Thus our hypothesis is now, so far as the χ^2 test is concerned, in agreement with experiment.

Experiments on the χ^2 Distribution.

22.30. Several statisticians have conducted experiments to verify the theory which we have discussed in the foregoing sections. A certain amount of work in this field remains to be done, but generally it may be said that experiment supports the theory. So far as cases where the m's are calculated *a priori* are concerned there is little doubt of its correctness.

In one set of experiments (ref. (511)) 200 beans were thrown into a revolving circular tray with 16 equal radial compartments and the number of beans falling into each compartment was counted. The 16 frequencies so obtained were arranged (1) in a 4 × 4 table, and (2) in a 2 × 8 table. χ^2 was calculated from the independence frequencies, as in Example 22.5.

The experiment and the calculations were repeated 100 times. The following table exhibits the actual and the theoretical distribution of χ^2:—

TABLE 22.6.—*Theoretical Distribution of χ^2, calculated from Independence Values, in Tables with 16 Compartments, compared with the Actual Distributions given by 100 Experimental Tables. In the first case ν must be taken as 9, in the second as 7.*

χ^2	4 Rows, 4 Columns.		2 Rows, 8 Columns.	
	Expectation.	Observation.	Expectation.	Observation.
0– 5	16·6	17	34·0	29·5
5–10	48·4	44	47·1	56·5
10–15	26·0	32	15·3	10
15–20	7·3	6	3·0	3
20–	1·8	1	0·6	1
Total	100·1	100	100·0	100

In a second experiment with 2 × 2 tables 350 experimental tables of 100 observations each were available. Table 22.7 shows the actual and theoretical distributions in this case.

TABLE 22.7.—*Theoretical Distribution of χ^2 for a Table with 2 Rows and 2 Columns, when χ^2 is calculated from the Independence Values, compared with the Actual Results for 350 Experimental Tables.*

Value of χ^2.	Number of Tables.	
	Expected.	Observed.
0 –0·25	134·02	122
0·25–0·50	48·15	54
0·50–0·75	32·56	41
0·75–1·00	24·21	24
1 –2	56·00	62
2 –3	25·91	18
3 –4	13·22	13
4 –5	7·05	6
5 –6	3·86	5
6–	5·01	5
Total . .	349·99	350

It is interesting to see what happens if we apply the χ^2 test to these tables.

In Table 22.6, grouping together the frequencies from $\chi^2 = 15$ upwards, so that $\nu = 3$, χ^2 is found to be 2·27 for the 4×4 tables and 4·36 for the 2×8 tables, giving $P = 0.52$ in the first case and 0·22 in the second.

In Table 22.7, $\chi^2 = 7.53$, $\nu = 9$, $P = 0.58$.

Goodness of Fit.

22.31. The χ^2 distribution, as we have seen, leads to tests of the correspondence between theory and fact, and this and other reasons have led to its being described as a test of the " goodness of fit." This expression may be used in two ways. In the first place, it may describe the " fit " of observed and hypothetical data. In the second, it may be used without reference to a hypothesis merely to provide an objective method of estimating the merits of a particular formula or a particular curve in graduating a set of values or a series of points.

The arithmetic in the second class of cases is exactly the same as in the first. Conventionally, we regard very low values of P as denoting a poor fit, and moderate values as denoting a reasonably good fit. High values show an excellent fit, and in considering them we take no heed of the point discussed in 22.19 (b), since we are assessing the closeness of the curve to the data, not the probability that the first represents a universe from which the second was derived by random sampling.

SUMMARY.

1.
$$\chi^2 = S\left\{\frac{(\tilde{m} - m)^2}{m}\right\}$$

$$= S\left(\frac{\tilde{m}^2}{m}\right) - N$$

where $\tilde{m}$ refers to the observed and m to the theoretical frequencies.

2. The number of degrees of freedom of an aggregate of cells is denoted by ν, and is equal to the number of cells whose frequencies can be determined at will. When ν cell frequencies are determined, the remainder are calculable directly from the conditions to which the cell frequencies are subjected by the nature of the data.

3. The frequency-distribution of χ^2 is given by

$$y = y_0 e^{-\frac{\chi^2}{2}} \chi^{\nu-1}$$

4. From this it is possible to ascertain the probability P that on random sampling we should get a value of χ^2 as great as or greater than a given value. Tables have been constructed for this purpose.

5. The χ^2 distribution may be applied to data grouped in cells provided (a) that the total number N in the sample is large, (b) that no theoretical cell frequency is small, and (c) that the constraints are linear.

6. The value of P for any given case enables us to judge of the correspondence between hypothesis and data.

7. When the theoretical cell frequencies have to be calculated from parameters estimated from the data, the χ^2 test can be applied with

$$\chi'^2 = S\frac{(\tilde{m} - m')^2}{m'}$$

instead of χ^2, provided that the cell frequencies are large, the estimates are "efficient," and the number of degrees of freedom used in ascertaining P is reduced by unity for every parameter which is estimated.

8. The value of P can also be used to give an objective criterion of the "goodness of fit" of a curve to a set of points or of a formula to a set of values.

EXERCISES.

22.1. The following table (Weldon) gives the results of a dice-throwing experiment:—

12 Dice thrown 4096 Times, a Throw of 6 reckoned a Success.

Number of Successes .	0	1	2	3	4	5	6	7 and over	Total.
Frequency .	447	1145	1181	796	380	115	24	8	4096

Find χ^2 on the hypothesis that the dice were unbiased and hence show that the data are consistent with this hypothesis so far as the χ^2 test is concerned.

22.2. Perform an experiment by throwing a die 600 times and noting the number of points at each throw. Use these data to inquire whether the die is biased.

22.3. 200 digits were chosen at random from a set of tables. The frequencies of the digits were:

Digit . . .	0	1	2	3	4	5	6	7	8	9	Total.
Frequency . .	18	19	23	21	16	25	22	20	21	15	200

Use the χ^2 test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the tables from which these were chosen.

22.4. Perform an experiment on the lines of Exercise 22.3 by taking, say, the last figure in 200 logarithms taken from a set of five-figure logarithm tables.

22.5. (Data: Yule, ref. (93).) Sixteen pieces of photographic paper were printed down to different depths of colour from nearly white to a very deep blackish brown. Small scraps were cut from each sheet and pasted on cards, two scraps on each card one above the other, combining scraps from the several sheets in all possible ways, so that there were 256 cards in the pack. Twenty observers then went through the pack independently, each one naming each tint either "light," "medium" or "dark."

The following table shows the name assigned to each of the two pieces of paper :—

Name assigned to Lower Tint.	Name assigned to Upper Tint.			Total.
	Light.	Medium.	Dark.	
Light . . .	850	571	580	2001
Medium. . .	618	593	455	1666
Dark . . .	540	456	457	1453
Total .	2008	1620	1492	5120

Show that there is a significant association between the name assigned to one piece and the name assigned to the other.

22.6. Apply the χ^2 test to the data of Example 3.9, page 44, and examine the justification for the conclusions there drawn.

22.7. Show that, if ν is large, P is below the 5 per cent. level of significance if

$$\sqrt{2\chi^2} - \sqrt{2\nu - 1} > 1.65$$

and below the 1 per cent. level of significance if

$$\sqrt{2\chi^2} - \sqrt{2\nu - 1} > 2.33$$

22.8. Table 5.6, page 78, gives the number of criminals of normal and weak intellect for various ranges of weight.

Assuming this to be a random sample of criminals, do the data support the suggestion that weak-minded criminals are not underweight?

22.9. Show that in a 2×2 contingency table wherein the frequencies are $\dfrac{a \mid b}{c \mid d}$, χ^2 calculated from the "independence" frequencies is

$$\frac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(b+d)(a+c)}$$

22.10. Show similarly that for a $2 \times n$ table

$$\chi^2 = S_r \left\{ \frac{N_1 N_2 \left(\frac{\mu_{1r}}{N_1} - \frac{\mu_{2r}}{N_2} \right)^2}{\mu_{1r} + \mu_{2r}} \right\}$$

where μ_{1r}, μ_{2r} are the 2 frequencies in the rth column and N_1, N_2 are the marginal sums of the 2 rows.

22.11. Two investigators draw samples from the same town in order to estimate the number of persons falling in the income groups "poorer," "middle class," "well to do." (The limits of the groups are defined in terms of money and are the same for both investigators.) Their results are as follows:—

Investigator.	Income Group.			
	"Poorer."	"Middle Class."	"Well to do."	Totals.
A	140	100	15	255
B	140	50	20	210
Totals	280	150	35	465

Show that the sampling technique of at least one of the investigators is suspect.

22.12. Exercise 10.17 gives the number of deaths per day of women over 85 published in *The Times* during 1910–12. Using the theoretical frequencies obtained in that exercise on the hypothesis that the numbers are distributed in a Poisson series, employ the χ^2 test to estimate the correctness of this hypothesis.

22.13. Design and execute an experiment involving the χ^2 test to test the randomicity of Tippett's numbers.

22.14. (Data: G. Mendel's classical paper on "Experiments in Plant-Hybridisation"—quoted in translation in W. Bateson's "*Mendel's Principles of Heredity*.")

In experiments on pea-breeding, Mendel obtained the following frequencies of seeds: 315 round and yellow; 101 wrinkled and yellow; 108 round and green; 32 wrinkled and green. Total, 556.

Theory predicts that the frequencies should be in the proportions $9 : 3 : 3 : 1$.

Examine the correspondence between theory and experiment, calculating P either directly (page 418, footnote) or by interpolation from tables.

22.15. A particular experiment gives, on hypothesis H, $\chi^2 = 9$, $\nu = 8$; when repeated it gives the same result. Show that the two results taken together do not give the same confidence in H as either taken separately.

CHAPTER 23.

THE SAMPLING OF VARIABLES—SMALL SAMPLES.

The Problem.

23.1. We now proceed to examine the theory of samples which are not large enough to warrant the assumptions underlying the work of Chapters 19 to 21. In particular, it will no longer be open to us to assume (*a*) that the random sampling distribution of a parameter is approximately normal, or even single-humped, or (*b*) that values given by the data are sufficiently close to the universe values for us to be able to use them in gauging the precision of our estimates.

The removal of these assumptions imposes severe restriction on our work, and, as we shall see, an entirely new technique is necessary to deal with the problems for which they are not permissible. The division between the theories of large and small samples is therefore a very real one, though it is not always easy to draw a precise line of demarcation. We should point out, however, that as a rule the methods of the theory of small samples are applicable to large samples, though the reverse is not true.

Estimates.

23.2. In the theory of large samples we were able to take the value of a parameter in a sample to be an estimate of that parameter in the universe. This procedure, obvious though it seems, is not in general valid for small samples. We must therefore discuss briefly the basis on which estimates of given parameters are to be made.

A full investigation of this question would take us far beyond the limits of this book. It involves matters of considerable mathematical and philosophical complexity, some of which still form the subject of dispute among statisticians. But in the theory of small samples the main parameters of interest are the mean and the standard deviation (or the variance), and we will proceed to consider these two.

Estimates of the Arithmetic Mean.

23.3. We shall take as the estimate of the arithmetic mean the value of the sample mean. That is to say, if we have n sample values x_1, x_2, . . . x_n, our estimate $\bar{x}$ of the mean in the universe is

$$\bar{x} = \frac{1}{n} S(x) \qquad . \qquad . \qquad . \qquad (23.1)$$

For estimates of the mean, therefore, the practice is the same for small samples as for large.

It may be shown that for samples from a normal universe an estimate

obtained in this way is the " best " in the sense that its sampling variance is less than that of any other estimate of the mean.

Estimates of the Variance.

23.4. Let us denote the variance in the universe by $\sigma_u{}^2$ and the mean by m.

If m is known, we take as an estimate of the variance the mean square deviation of the sample about m; *i.e.* the estimate, which we write as $\sigma_s{}^2$, is given by

$$\sigma_s{}^2 = \frac{1}{n}S(x-m)^2 \quad . \qquad . \qquad . \qquad . \quad (23.2)$$

In general, however, we do not know the value of m, which will itself have to be estimated. In this case equation (23.2) is no longer applicable.

23.5. If m is the universe mean and $\bar{x}$ is the sample mean, we have :

$$S(x-m)^2 = S(x - \bar{x} + \bar{x} - m)^2$$
$$= S(x - \bar{x})^2 + S(\bar{x} - m)^2$$
$$= S(x - \bar{x})^2 + n(\bar{x} - m)^2$$

Hence,

$$\sigma_s{}^2 = \frac{1}{n}S(x-\bar{x})^2 + (\bar{x} - m)^2 \quad .$$

The term $\frac{1}{n}S(x-\bar{x})^2$ is the variance of the sample. We see that it differs from $\sigma_s{}^2$ by the term $(\bar{x} - m)^2$.

Now this term will not, in general, vanish ; nor will it vanish on the average in a large number of cases, for it is essentially positive. Hence, if we take the variance of the sample to be an estimate of the variance of the universe we shall involve ourselves in a systematic error of magnitude $(\bar{x} - m)^2$.

This term is the square of the deviation of the mean of the sample from the mean of the universe, and its average value in a large number of samples is the variance of the mean, which we know to be equal to $\frac{\sigma_u{}^2}{n}$.

It seems reasonable, therefore, instead of ignoring the presence of the term $(\bar{x} - m)^2$, to take it as equal to $\frac{\sigma_u{}^2}{n}$. We will attempt, on this basis, a new estimate, which we shall write $_c\sigma_s{}^2$. We have then :

$$_c\sigma_s{}^2 = \frac{1}{n}S(x-\bar{x})^2 + \frac{\sigma_u{}^2}{n}$$

The value of σ_u is unknown, but we may, as an approximation, write $_c\sigma_s$ instead. If we do so we get :

$$_c\sigma_s{}^2 = \frac{1}{n}S(x-\bar{x})^2 + \frac{1}{n}{}_c\sigma_s{}^2$$

$$_c\sigma_s{}^2 = \frac{1}{n-1}S(x-\bar{x})^2 \quad . \qquad . \qquad . \quad (23.3)$$

The effect of taking $_c\sigma_s{}^2$ given by equation (23.3), instead of the variance of the sample, will thus be to eliminate the systematic error of estimation to which we have just referred.

23.6. We may look at this in a slightly different way. Suppose we take a large number of estimates of the variance of a universe compiled according to equation (23.2), m being assumed known. These estimates will fall into a distribution which is the sampling distribution of the variance in samples of n. If, as will usually be the case, it is of the single-humped type, we expect it to have a mean located at the true value of the variance in the universe.

Now if we take as estimates of the variance the variance of the samples (each about its own sample mean), the above will not be true, owing to the small systematic shift represented by the term $(\bar{x} - m)^2$; but it will be true of the estimates given by equation (23.3), and this is therefore a preferable estimate to take.

23.7. Equation (23.3) was obtained by reasoning which does not depend on the size of n, and strictly speaking we should take it as applicable also to large samples. But if n is large, n and $n - 1$ are for all practical purposes equal. With such samples our results are true only within the range of the standard error, which is usually of order $\dfrac{1}{\sqrt{n}}$, and there is little point in straining after an illusory refinement by taking $n - 1$ instead of n in calculating the variance.

From a similar point of view it might be thought that since the term $\dfrac{\sigma_u{}^2}{n}$ is generally less than the square of the standard error of the variance, it is equally idle to make allowance for it in estimating the variance. This would be true if the term were zero on the average; but in fact it is not, being a biased error, and we are justified in the long run in allowing for it.

Furthermore, we may point out that the use of $_c\sigma_s{}^2$, the corrected value obtained by allowing for the term $\dfrac{\sigma_u{}^2}{n}$, is only valid *on the average.* If, on random sampling, we get a sample variance greater than the universe variance, the correction only makes matters worse, and may even lead to an absurd result. An instance happens to occur in **23.33** below.

Degrees of Freedom of an Estimate.

23.8. In discussing the χ^2 test we introduced the notion of **number of degrees of freedom,** being the number of cells in an aggregate whose frequency could be assigned at will. We may conveniently extend this nomenclature to estimates of parameters and particularly of variance.

We shall refer to the divisor in the estimates of equations (23.1), (23.2) and (23.3) as the number of degrees of freedom of the estimates, and shall write it as ν. Thus, ν in equation (23.2) is n, and in equation (23.3) is $n - 1$.

That this convention conforms to that adopted for the χ^2 test may easily be seen. We saw that ν is the number of cells, that is, the number of terms contributing to the χ^2 sum, less one for each constraint and one for each parameter which had been estimated from the data. In the

quantity $S(x-m)^2$ there are n independent contributions of the type $(x-m)^2$, and hence we may say that n is the number of degrees of freedom of that estimate ; but in the quantity $S(x-\bar{x})^2$ we have used the data to estimate $\bar{x}$, and hence the number of degrees of freedom is lowered by unity, *i.e.* equals $n-1$.

Tests of Significance.

23.9. It cannot be over-emphasised that estimates from small samples are of little value in indicating the true value of the parameter which is estimated. Some estimates will be better than others, but no estimate is very reliable. In the present state of our knowledge this is particularly true of samples from universes which are suspected not to be normal.

Nevertheless, circumstances sometimes drive us to base inferences, however tentatively, on scanty data. In such cases we can rarely, if ever, make any confident attempt at locating the value of a parameter within serviceably narrow limits. For this reason we are usually concerned, in the theory of small samples, not with estimating the actual value of a parameter, but in ascertaining whether observed values can have arisen by sampling fluctuations from some value given in advance. For example, if a sample of ten gives a correlation coefficient of $+0\cdot1$, we shall inquire, not the value of the correlation in the parent universe, but, more generally, whether this value can have arisen from an uncorrelated universe, *i.e.* whether it is *significant* of correlation in the parent.

23.10. The remainder of this chapter will accordingly be devoted to a brief discussion of various **tests of significance**. Within this book we shall not have space to deal with these tests as fully as we should like ; but our account of sampling methods would be incomplete without some reference to sundry results of great intrinsic interest and importance in the field of small samples.

The Assumption of Normality.

23.11. We have already considered one test of significance, that given by the distribution of χ^2. This is one of the simplest and most general tests known ; but the student will recall that it depends on the assumption that the theoretical distribution of cell frequencies in each cell is normal. This is justified under the conditions laid down in **22.18.**

In the tests which we shall now discuss we are similarly compelled to make some assumption about the nature of the parent universe, although we shall no longer be able to lay down analogous conditions on the arrangement of the data under which the assumption is justified. We shall specifically assume that the parent universe is normal unless otherwise stated.

23.12. Our results will, therefore, be strictly true only for the normal universe. Some experiments have been made to throw light on the question whether they are true for other types of universe. It appears that, provided the divergence of the parent from normality is not too great, the results which are given below as true for normal universes are true to a large extent for other universes. But the whole situation is obscure, and it is to be hoped that in time investigators will be able to engage in the labour of a closer inquiry. In any case, if there is any good reason to

suspect that the parent is markedly skew, *e.g.* U- or J-shaped, the methods of the succeeding sections cannot be applied with any confidence.

23.13.   We may direct attention to one further point on which caution is necessary.   In the theory of large samples we recommended the student to base his conclusions on a range of six times the standard error, and pointed out that for normal universes the probability of deviations from the true value outside this range was less than 3 in 1000.   One can feel great confidence in conclusions supported by probabilities of this order. But in the theory of small samples it is, as a rule, necessary to use larger probabilities, say, of one in 20 or one in 100, *e.g.* the 1 per cent. and 5 per cent. levels of P in the χ^2 test.   The force of inferences based on probabilities of this order is not so great as before, and the student should bear this fact in mind.

23.14.   For a known parent universe, and in particular for a normal parent, it is not difficult to find expressions for the random sampling distribution of the commoner parameters such as the mean and standard deviation.   But these distributions, even when mathematically tractable, will in general contain certain parent values.   For instance, the sampling distribution of the means of samples of n from a normal universe with mean m and standard deviation σ is also normal with mean m and standard deviation $\dfrac{\sigma}{\sqrt{n}}$.   In the cases which we wish to consider, n is not large enough for us to take estimates of m and σ from the sample to find the sampling distribution to any close degree of approximation.

It is, however, a remarkable fact that we can construct certain parameters whose sampling distributions are either independent of, or dependent on only one of, the constants of the parent.   We will proceed to consider two important distributions of this kind, the so-called t-distribution, due to "Student," and the z-distribution, due to R. A. Fisher.

The t-Distribution.

23.15.   Writing, as before,

$$\bar{x} = \frac{1}{n}S(x)$$

$$_c\sigma_s{}^2 = \frac{1}{n-1}S(x - \bar{x})^2$$

let us define a new parameter t by the equation

$$t = \frac{\bar{x} - m}{_c\sigma_s}\sqrt{\nu + 1} \qquad . \qquad . \qquad . \qquad . \quad (23.4)$$

where $\nu = n - 1$ and m is the mean of the universe.

We shall refer to ν as the number of degrees of freedom of t.

Then it may be shown that, for samples of n from a normal population, the distribution of t is given by

$$y = \frac{y_0}{\left(1 + \dfrac{t^2}{\nu}\right)^{\frac{\nu + 1}{2}}} \qquad . \qquad . \qquad . \qquad . \quad (23.5)$$

23.16. We will imagine y_0 chosen so that the area of the curve given by equation (23.5) is unity. Then, precisely as for the χ^2 distribution, the probability P_s that, on random sampling, we shall get a value of t not greater than some value t_0 is the area of the curve to the left of the ordinate at the point t_0. We may write this

$$P_s = \int_{-\infty}^{t_0} \frac{y_0 dt}{\left(1 + \frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}} \qquad . \qquad . \qquad . \qquad (23.6)$$

Similarly, the probability that we get a value of t between the limits t_1 and t_2 is given by

$$P_s = \int_{t_1}^{t_2} \frac{y_0 dt}{\left(1 + \frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}} \qquad . \qquad . \qquad . \qquad (23.7)$$

Form of " Student's " Distribution.

23.17. The curves given by equation (23.5) are easy to study. Clearly they are symmetrical about $t = 0$, since only even powers of t appear in their equation. Further, since $\dfrac{1}{\left(1 + \dfrac{t^2}{\nu}\right)}$ decreases as t increases, the curves will have a mode (coinciding, of course, with the mean) at $t = 0$, and will tail off to infinity on each side. They will, in fact, be symmetrical single-humped curves rather like the normal curve, only more leptokurtic.

As ν tends to infinity, $\dfrac{1}{\left(1 + \dfrac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}}$ tends to $e^{-\frac{t^2}{2}}$, and hence t is distributed normally. This fact enables us to use the tables of the normal integral to evaluate P approximately when ν is large.

23.18. At the end of this book we reproduce by permission tables of the integral (23.6) calculated by " Student " himself (Appendix Table 5). These have been reduced to three places of decimals from the original four.

Tables of rather a different form have been given in " *Tables for Statisticians and Biometricians, Part I,*" and by R. A. Fisher, and to avoid possible confusion we point out where these tables differ.

" *Tables for Statisticians, etc.,*" gives the values of

$$P_T = \int_{-\infty}^{z_0} \frac{y_0 dz}{(1 + z^2)^{\frac{\nu+1}{2}}}$$

where $z = \dfrac{t}{\sqrt{\nu}}$, for ν from 1 to 9. These values (which were also calculated by " Student ") are of the same kind as, but more limited in range than, those of our table.

R. A. Fisher, in his " *Statistical Methods for Research Workers,*" adopts the standpoint we have already noticed in discussing the χ^2 distribution

(Chapter 22), and gives values of t corresponding to various values of ν and the 5 per cent. and 1 per cent. levels of a third probability P_F.

P_s and P_F are simply related. P_s is the probability that an observed value will *not* exceed t_0. P_F is the probability that an observed value of t, *regardless of sign*, will exceed t_0.

Hence,

$P_s =$ Area of curve to the left of ordinate t_0

$P_F =$ Area to right of t_0 + area to left of $-t_0$

$= 2$ (Area to right of t_0) (since the curve is symmetrical)

$= 2\,(1 - P_s)$ . . . . (23.8)

The student should keep these relations in mind, particularly when thinking of levels of significance. In Fisher's sense a value of P_F will fall below the 5 per cent. level if P_F is less than 0·05. This implies that P_s is greater than 0·975, not 0·95.[1]

Applications of " Student's " Distribution.

23.19. We proceed to give one or two examples of the way in which the " Student " distribution is generally used to test the significance of various results obtained from small samples.

Example 23.1.—Ten individuals are chosen at random from a population and their heights are found to be, in inches, 63, 63, 66, 67, 68, 69, 70, 70, 71 and 71. In the light of these data, to discuss the suggestion that the mean height in the universe is 66 inches.

In the first place, let us note that the universe is likely to be approximately normal, from our knowledge of height distributions, and the sampling is random.

In the sample we find that

$$\bar{x} = 67 \cdot 8 \text{ inches}$$

and

$$_c\sigma_s = 3 \cdot 011 \text{ inches}$$

Let us now calculate t from equation (23.4), taking m to be 66 inches. We have :

$$t = \frac{67 \cdot 8 - 66}{3 \cdot 011}\sqrt{10} = 1 \cdot 89$$

From the Appendix Table 5 (column $\nu = 9$):

for $t = 1 \cdot 8$, $P = 0 \cdot 947$

for $t = 1 \cdot 9$, $P = 0 \cdot 955$

Hence,

for $t = 1 \cdot 89$, $P = 0 \cdot 954$

---

[1] A comparison of the tables is not made any easier by the fact that "Student" and Fisher use n to denote the degrees of freedom, whereas " *Tables for Statisticians* " uses it to denote the number in the sample. We noted the same conflict in the χ^2 tables. We hope here that the use of a separate symbol ν will remove a good deal of the confusion.

The distinction between P_S and P_F did not arise in Chapter 22 because χ^2 is essentially positive.

Thus the chance of getting a value of t greater than that observed is $1 - 0.954$, *i.e.* 0.046, or about one in twenty. The probability of getting t greater in *absolute value* is 0.092, or about one in ten. We should hardly regard this as significant; but if we did, we should argue that as the observed value of t is improbable, the initial assumptions on which we obtained it were incorrect; and this in turn suggests that there is some doubt about the true mean being 66 inches.

Example 23.2.—(Voelcker's data quoted by " Student," *Biometrika*, vol. 6, 1908-9, p. 19.)

Voelcker grew certain crops of potatoes dressed (*a*) with sulphate of potash, and (*b*) with kainite. In four experiments, two of each of 1904 and 1905, the differences in yields per acre (sulphate plot less kainite plot) were:

$$0.5464 \text{ ton}$$
$$0.3013 \text{ ,,}$$
$$1.5241 \text{ ,,}$$
$$0.6786 \text{ ,,}$$

This suggests that sulphate of potash is a better manure than kainite. Required to discuss the question.

From our knowledge of crop yields we expect them to be distributed in a single-humped form not very far removed from the normal. Let us suppose that the two manures have the same effect on yield. Then the differences of plots will be distributed in an approximately normal form about zero mean.

The mean of the four differences is 0.7626 ton, and we find $_c\sigma_s = 0.5312$. Hence,

$$t = \frac{0.7626 - 0}{0.5312}\sqrt{4}$$

$$= 2.871$$

From the tables, for $\nu = 3$, $P = 0.968$ approximately.

Hence the chance P of getting a value of t greater than that observed is about 1 in 33. The chance of getting a value greater *absolutely* than the observed value is 0.06. If we choose to regard this as significant, we are led to suspect our hypothesis that the two manures exert equal influences on yield, and hence to suppose, though with little confidence so far as these data are concerned, that sulphate of potash is the better manure.

23.20. The student who wishes to apply the t-distribution for himself is advised to make a careful study of the logic of the argument underlying the inferences we have drawn in the foregoing two examples.

In Example 23.1 we saw that the chance of getting a value of t less than 1.89 is approximately 0.954. This is not the same thing as saying that the probability of a deviation in the sample mean of 1.8 inches or less is 0.954. In fact, we do not know this probability, and the smallness of the sample prevents us from approximating to it with any closeness.

It *might* happen that σ in the universe was such that a deviation of 1·8 inches was not at all improbable. The relative improbability of t would then be due to deviations of $_c\sigma_s$ from σ_u.

Comparison of Two Samples.

23.21. Suppose we have two samples $x_1,\ x_2,\ \ldots\ x_{n_1}$ and x_1', $x_2',\ \ldots\ x_{n_2}'$. Let us, as before, define

$$\left.\begin{array}{l} \bar{x}_1 = \dfrac{1}{n_1}\,\mathrm{S}(x) \\[2mm] \bar{x}_2 = \dfrac{1}{n_2}\,\mathrm{S}(x') \\[2mm] _c\sigma_{s_1}^2 = \dfrac{1}{n_1 - 1}\,\mathrm{S}(x - \bar{x}_1)^2 \\[2mm] _c\sigma_{s_2}^2 = \dfrac{1}{n_2 - 1}\,\mathrm{S}(x' - \bar{x}_2)^2 \end{array}\right\} \qquad . \qquad . \qquad . \text{(23.9)}$$

Let us further define

$$_c\sigma_s{}^2 = \frac{1}{n_1 + n_2 - 2}\{\mathrm{S}(x - \bar{x}_1)^2 + \mathrm{S}(x' - \bar{x}_2)^2\} \qquad . \qquad . \text{(23.10)}$$

If the two samples come from the same universe, $_c\sigma_s{}^2$ will be an estimate of $\sigma_u{}^2$. It has, as we might expect, $n_1 + n_2 - 2$ degrees of freedom, since both $\bar{x}_1$ and $\bar{x}_2$ are calculated from the data.

Let us write

$$\nu = n_1 + n_2 - 2 \qquad . \qquad . \qquad . \qquad . \text{(23.11)}$$

and define

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{_c\sigma_s \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \\[3mm] &= \frac{\bar{x}_1 - \bar{x}_2}{_c\sigma_s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\ . \qquad . \qquad . \qquad . \text{(23.12)} \end{aligned}$$

Then it may be shown that t, as so defined, is distributed according to the form of equation (23.5) with ν degrees of freedom.

Example 23.3.—(Data from R. A. Fisher, *Metron*, vol. 5, 1925, p. 95.)

Eight pots growing three barley plants each were exposed to a high tension discharge, while nine similar pots were enclosed in an earthed wire cage. The numbers of tillers in each pot were as follows :—

Caged .	.	.	17, 27, 18, 25, 27, 29, 27, 23, 17
Electrified .	.	.	16, 16, 20, 16, 20, 17, 15, 21

We are interested in the question whether electrification exercises any real effect on the tillering.

We find

$$\bar{x}_1 = 23 \cdot 338 \qquad \bar{x}_2 = 17 \cdot 625$$

$$\bar{x}_1 - \bar{x}_2 = 5 \cdot 708$$

$$_c\sigma_s{}^2 = \frac{1}{15} 221 \cdot 875 = 14 \cdot 7916 \qquad _c\sigma_s = 3 \cdot 846$$

$$t = \frac{5 \cdot 708}{3 \cdot 846} \sqrt{\frac{8 \times 9}{17}} = 3 \cdot 05$$

$$\nu = 8 + 9 - 2 = 15$$

From the tables we find that $P_s = 0 \cdot 996$.

Hence, if the samples came from the same universe, they furnish a value of t which is improbable—an absolutely greater value would arise only 8 times in a thousand. We therefore suspect that the universes are different, *i.e.* that electrification does exert some effect on the tillering.

23.22. In applying the t-distribution to two samples as in the preceding example one further point should be borne in mind. It does not follow from a significant value of t that the samples come from universes which have different means. Samples from two universes with the same means and different standard deviations would also furnish significant t's on occasion. We can test whether this is so by the method of **23.24** below.

Significance of Regression Coefficients.

23.23. R. A. Fisher has shown that the " Student " distribution can be applied to test the significance of regression coefficients and also of certain curvilinear regressions. We have not the space here to give a discussion of these results, but the reader is referred to ref. (536) for further particulars. A test of the significance of correlation coefficients is given below (**23.34** to **23.39**).

Fisher's z-Distribution.

23.24. Suppose that we have two samples, as in **23.21**, with estimated variances $_c\sigma_{s_1}^2$ and $_c\sigma_{s_2}^2$ as defined in equation (23.9).

Put

$$z = \tfrac{1}{2} \log_e \frac{_c\sigma_{s_1}^2}{_c\sigma_{s_2}^2} \qquad . \qquad . \qquad . \qquad . \qquad (23.13)$$

and write

$$\left. \begin{array}{c} \nu_1 = n_1 - 1 \\ \nu_2 = n_2 - 1 \end{array} \right\} \qquad . \qquad . \qquad . \qquad (23.14)$$

so that ν_1 and ν_2 are the degrees of freedom of the estimates $_c\sigma_{s_1}^2$ and $_c\sigma_{s_2}^2$.

Then R. A. Fisher has shown that, if the samples come from the same universe and that universe is normal, z is distributed according to the law

$$y = y_0 \frac{e^{\nu_1 z}}{\left(\nu_1 e^{2z} + \nu_2 \right)^{\frac{\nu_1 + \nu_2}{2}}} \qquad . \qquad . \qquad . \qquad (23.15)$$

As usual, we take y_0 so that the area of the curve is unity, and the probability that we get a given value z_0 or greater on random sampling will be given by the area to the right of the ordinate at z_0.

23.25. This probability is not easy to tabulate owing to the fact that it depends upon the two numbers ν_1 and ν_2. Fisher has therefore prepared tables showing the 5 per cent. and 1 per cent. significance points of z, and a further table of the 0·1 per cent. points has been given by Colcord and Deming. These tables are reproduced by permission in Appendix Tables 6A, 6B and 6C. For practical purposes they are sufficient to enable the significance of an observed value of z to be gauged. If the exact value of the probability of obtaining a given value of z or greater is required, use may sometimes be made of the tables of the incomplete beta-function (ref. (600)).

Example 23.4.—Consider again the data of Example 23.3.

Here, as always, it is convenient to take the suffix 1 to refer to the larger of the two estimates of variance.

We have :

$$_c\sigma_{s_1}^2 = \frac{184}{8} = 23$$

$$_c\sigma_{s_2}^2 = \frac{37 \cdot 875}{7} = 5 \cdot 4107$$

$$z = \tfrac{1}{2} \log_e \frac{23}{5 \cdot 4107}$$

$$= 0 \cdot 724$$

$$\nu_1 = 8, \qquad \nu_2 = 7$$

From Appendix Table 6A we see that for these degrees of freedom the 5 per cent. significance value of z is 0·6576. From Table 6B the 1 per cent. value is 0·9614.

The observed z lies between these two and is thus of rather doubtful significance.

The Analysis of Variance.

23.26. This is the name given to a process now frequently applied, mainly in agricultural experiments. For a full treatment we must refer the reader to those works dealing with the latter subject ; here we will do no more than attempt to explain the general principles of the method.

Suppose we have n varieties of barley and desire to determine whether they differ significantly in yield per acre. It would be no good growing just one plot of each and comparing yields, for soil is very variable and we should have no idea whether any observed differences in yield were due to differences in variety or to differences in soil or some other such factor.

Let us then grow k plots of the same size for each variety. We shall then have data to determine the standard error of the mean yield for each variety and so the standard error of each difference of mean yield. But the process may be simplified. If we scatter the plots well in amongst one another, preferably at random, we may expect that fluctuations in soil from plot to plot will affect all varieties to about the same extent, and

consequently the standard deviations of the varieties will not differ significantly owing to soil influences.

Let $\sigma_1, \sigma_2, \ldots \sigma_p, \ldots \sigma_n$ be the standard deviations of the yields of the several varieties and

$$\sigma_v^2 = \frac{S(\sigma_p^2)}{n} \qquad . \qquad . \qquad . \qquad (23.16)$$

Supposing for simplicity that n is large enough for us to be able to ignore the correction of equation (23.3), we may, on the hypothesis that the yields of different varieties are equal, take σ_v^2 to be an estimate of the value of the variance of a variety.

Also, if $\bar{x}$ be the general mean of all yields and $\bar{x}_1, \bar{x}_2, \ldots \bar{x}_p, \ldots \bar{x}_n$ be the means of the several varieties, the variance of the means is given by

$$\sigma_m^2 = \frac{S(\bar{x}_p - \bar{x})^2}{n} \qquad . \qquad . \qquad (23.17)$$

Now the variance of the distribution of means of samples of k is $\frac{\sigma_v^2}{k}$.

Hence, if

$$\sigma_m^2 > \frac{\sigma_v^2}{k}$$

or

$$k\sigma_m^2 > \sigma_v^2 \qquad . \qquad . \qquad . \qquad (23.18)$$

significantly, we may take it that the varieties do differ significantly in yield.

23.27. If σ_y^2 be the variance of yield of all the plots taken together without regard to variety, we have a simple relation between σ_y^2, σ_m^2 and σ_v^2.

In fact, for any one variety, the sum of squares of deviations from the *general* mean is

$$k\{\sigma_p^2 + (\bar{x}_p - \bar{x})^2\}$$

and hence, summing for all varieties and dividing by nk, we have:

$$\sigma_y^2 = \sigma_v^2 + \sigma_m^2 \qquad . \qquad . \qquad . \qquad (23.19)$$

In this way we have *analysed the variance* of the total into two components, the variance of the means and the variance within the varieties.

23.28. It is convenient to arrange the results we have just obtained in the form of a table. The student will have no difficulty in recognising that, although we have talked of plots of barley to fix the ideas, similar analysis applies to any data in which we have n classes each of k members.

Since we want finally to compare σ_v^2 with $k\sigma_m^2$, and not with σ_m^2, it will be more convenient to put $k\sigma_m^2$ rather than σ_m^2 itself in a summary table (Table 23.1, page 446).

In the second sum of column 3, the summation is understood to relate to the squares of deviations of individuals from the mean of classes in which they occur, *i.e.* $\overset{r=nk}{\underset{r=1}{S}} (x_r - \bar{x}_p)^2$ is an abbreviation for

$$\overset{p=n}{\underset{p=1}{S}}\left\{\overset{r=k}{\underset{r=1}{S}} (x_{rp} - \bar{x}_p)^2\right\}$$

x_{rp} being the rth member of the pth class.

TABLE 23.1.

1.	2.	3.	4.	5.
Sums relating to Variation.	Divisor.	Sums.	Quotients.	
Between class means . .	n	$\underset{p=1}{\overset{p=n}{kS}} (\bar{x}_p - \bar{x})^2$	$k\sigma_m^2$	
Within classes . . .	nk	$\underset{r=1}{\overset{r=nk}{S}} (x_r - \bar{x}_p)^2$	σ_v^2	
Total . .	nk	$\underset{r=1}{\overset{r=nk}{S}} (x_r - \bar{x})^2$	σ_y^2	

As a check, we note that the first two items in column 3 must add up to the third. In actual practice it is customary to use this fact to deduce the second from the other two, and not work them out independently.

23.29. Let us take the following data as an illustration—an illustration only, for (1) n is not large, and (2) the data are a mere extract from an experiment on a much larger scale with 18, not 6, plots to each variety.

TABLE 23.2.—*Yield of Grain in grammes on Plots of Barley of One Square Yard, there being Five Varieties and Six Plots of Each.* (Data quoted by Engledow and Yule, "*The Principles and Practice of Yield Trials,*" 1926.)

(The tabular arrangement does not, of course, represent the physical lay-out of the plots.)

Plot Number.	Variety.					Mean.
	1	2	3	4	5	
1	387	372	350	340	398	369·4
2	420	455	417	360	358	402·0
3	353	375	400	358	334	364·0
4	331	328	325	370	340	338·8
5	358	383	378	395	320	366·8
6	400	308	275	375	430	357·6
Mean	374·8	370·2	357·5	366·3	363·3	366·4

The mean of the whole, $\bar{x}$, is 366·4. The sums of squares of deviations from this mean may be found in the usual way, and the calculation simplified by taking a working mean at, say, 366.

We find, to the nearest unit,

$$\mathop{S}_{r=1}^{r=nk} (x_r - \bar{x})^2 = 43,934$$

Similarly,

$$\mathop{S}_{p=1}^{p=n} (\bar{x}_p - \bar{x})^2 = 1,043$$

Hence the table of the analysis is as follows :—

TABLE 23.3.

1.	2.	3.	4.	5.
Sums relating to Variation.	Divisor.	Sums.	Quotients.	
Between class means. .	5	1,043	$k\sigma_m^2$	$= 209$
Within classes . . .	30	42,891	σ_v^2	$=1,430$
Total . .	30	43,934	σ_v^2	$=1,464$

We see that σ_v^2 is very much greater than $k\sigma_m^2$, and the magnitude of the difference suggests that it is due to some real cause.

We should probably infer that, since the·variability within a variety is greater than that between means of varieties, no significance can be attached to differences between the latter.

23.30. But the process of the previous section is not very accurate with samples so small as those with which we have been dealing. The corrected variances, based on degrees of freedom, not the number of observations, should be used (*cf.* equation (23.3)). This gives a more complex appearance to the arithmetic, but the principles are similar. The student will probably find the determination of the degrees of freedom his principal difficulty.

There are n class means, so that the number of degrees of freedom in the variance between class means is $n - 1$. There are k members in each class (degrees of freedom $k - 1$), and n classes, total $n(k - 1)$ degrees of freedom in the variance within varieties. For all classes together there are nk observations and hence $nk - 1$ degrees of freedom.

But

$$(nk - 1) = (n - 1) + n(k - 1)$$

and hence the degrees of freedom check by addition in the same way as the sums of squares.

23.31. Our general table now takes the form of Table 23.4, page 448, where we have used the symbols $_c\sigma_m^2$, $_c\sigma_v^2$, $_c\sigma_v^2$ to denote the variances corrected as in equation (23.3).

The student should note that these corrected variances are not additive. Nevertheless, it is common to refer to a process of analysis such as this as the " analysis of variance." Strictly speaking, perhaps, this is a misnomer. It is only the sum of column 3 which is analysed into component sums.

TABLE 23.4.

1.	2.	3.	4.	5.
Sums relating to Variation.	Divisor (Degrees of Freedom).	Sums of Squares.	Quotients.	
Between class means	$n-1$	$kS_{p=1}^{p=n} (\bar{x}_p - \bar{x})^2$	$k_c\sigma_m{}^2$	
Within classes .	$n(k-1)$	$S_{r=1}^{r=kn} (x_r - \bar{x}_p)^2$	$_c\sigma_v{}^2$	
Total	$nk-1$	$S_{r=1}^{r=kn} (x_r - \bar{x})^2$	$_c\sigma_v{}^2$	

23.32. In small samples the significance of the difference of $k_c\sigma_m{}^2$ and $_c\sigma_v{}^2$ can be ascertained by the z test, the appropriate degrees of freedom being those of column 2.

In fact, if the classes exercise no effect on the variate values of their members, so that the nk members can be regarded as a homogeneous set grouped at random into n classes, $k_c\sigma_m{}^2$ and $_c\sigma_v{}^2$ will be estimates of the variance in the universe. Further, if the parent universe is normal these estimates will be independent, for errors of estimation in the means of classes will be independent of errors in the variances within classes.[1] All the conditions for the application of the z test therefore obtain. If the test reveals no significance in the difference between $k_c\sigma_m{}^2$ and $_c\sigma_v{}^2$, we conclude that, so far as this approach shows, the class does not exert any distinguishing effect on its members. If, on the other hand, the difference is shown to be significant, the class does exert some influence.

Two cases may arise, according as $k_c\sigma_m{}^2$ is less than, or greater than, $_c\sigma_v{}^2$. It may be shown that these cases correspond to the existence of positive or negative intraclass correlation (**13.29**).

23.33. Table 23.3, with corrected variances, now becomes :

TABLE 23.5.

1.	2.	3.	4.	5.
Sums relating to Variation.	Degrees of Freedom.	Sums of Squares.	Quotients.	
Between class means	4	1,043	$k_c\sigma_m{}^2$	261
Within classes .	25	42,891	$_c\sigma_v{}^2$	1,716
Total	29	43,934	$_c\sigma_v{}^2$	1,515

---

[1] We proved on page 405 that for large samples errors in the mean and s.d. are uncorrelated in a symmetrical universe. It may be shown generally that for samples of any size from a normal universe the errors are *independent*.

We see at once that since the corrected variance within varieties is greater than that between varieties, any intraclass correlation must be negative.   To test its significance we have :

$$\nu_1 = 25 \qquad \nu_2 = 4$$

$$z = \tfrac{1}{2} \log_e \frac{1716}{261}$$

$$= 0 \cdot 942$$

From the Appendix Tables we see that the 5 per cent. point is about 0·876 and the 1 per cent. point 1·31. The result thus is barely significant.

It is instructive to note that the correction of the variances happens in this case to give an absurd result, such as was noted in 23.7 might occur ; the variance within classes is made to appear greater than the total variance, which is impossible.

Correlation Coefficient in Small Samples.

23.34. Although the distribution of the correlation coefficient in samples from a bivariate normal universe tends to the normal form as the size of the sample increases, a fact which justifies the use of the standard error for large n, the distribution diverges very remarkably from the normal when n is small, and even when n is moderately large if the correlation in the parent universe is high. Further investigation is therefore necessary before we can assess the significance of correlation coefficients obtained from small samples.

23.35. The distribution of the correlation coefficient in samples from a bivariate normal universe was obtained in an exact form by R. A. Fisher in 1915. Ordinates of the frequency-curves which give the distribution have been worked out for various values of n and ρ, the correlation in the universe, and are tabulated in "*Tables for Statisticians and Biometricians, Part 2,*" and more fully in ref. (577). The general form of these curves is illustrated in fig. 23.1, which shows the curves for $\rho = +0·6$ and various values of n.

A glance at this figure will show that even for a moderate value of ρ, such as $+0·6$, the distribution of the coefficient is U-shaped for $n = 3$, and, although single-humped, distinctly skew to the eye even for $n = 20$. For high values of ρ, such as $+0·9$, the distribution is skew for higher values of n.

As a result it is safe to say that the values of correlation coefficients calculated from samples of less than five will throw no light on the existence of correlation in the universe. For samples of 20 or 30 we cannot apply the standard error with much confidence if the correlation in the universe is likely to be very high, whether positive or negative. 50 seems to be the minimum number in the sample for the application of the standard error if ρ is very high, and 100 is safer.

23.36. Owing to the complexity of the equation which gives the distribution of the correlation coefficient, no tables have been published showing the areas of the frequency curves cut off by various ordinates.[1]

---

[1] Such tables are, however, promised from the Biometric Laboratory, University College, London, and should be published during 1937.

There are, therefore, no practical methods of assessing the reliability of an observed coefficient in small samples, such as we have been able to use for the normal curve, the χ^2-distribution, and the t- and z-distributions. We shall have to fall back on a procedure of transformation due to R. A. Fisher.
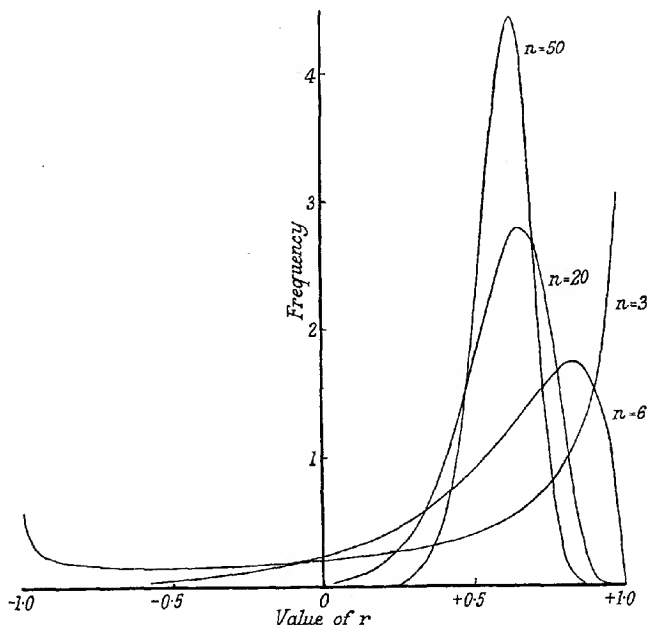
F IG. 23.1.—Frequency Distribution of the Correlation Coefficient in Samples from a Normal Universe with Correlation $+0\cdot6$ for Various Values of the Number in the Sample n. (In each case the total frequency, *i.e.* the area under the curve, is unity.)

23.37. Before we discuss this process, however, it is desirable to point out the degree of applicability of our results.

(1) In the first place, it has been shown that the distribution of partial correlation coefficients in samples of n is of the same form as that of total correlation coefficients in samples of $n-p$, where p is the number of *secondary* subscripts in the partial coefficient.

(2) Secondly, our results are strictly true only for normal universes. There is some experimental evidence to show that they are true for all practical purposes even if the parent is moderately skew but remains of the single-humped type ; but if there is any reason to suppose that the parent is J- or U-shaped according to one or more variates, the student should draw his conclusions with the utmost reserve.

Fisher's Transformation.

23.38.   If r and ρ are the correlations in the sample and the universe respectively, let us put

$$r = \tanh z \qquad \rho = \tanh \zeta$$

So that

$$z^* = \tfrac{1}{2} \log_e \frac{1+r}{1-r} \Bigg\}$$
$$\zeta = \tfrac{1}{2} \log_e \frac{1+\rho}{1-\rho} \Bigg\} \qquad . \qquad . \qquad . \qquad (23.20)$$

Then it may be shown that z is, to a close approximation, distributed normally about mean ζ with standard deviation $\dfrac{1}{\sqrt{n-3}}$.

In fact, the mean of z is given by

$$\bar{z} = \zeta + \frac{\rho}{2(n-1)} + \text{terms in } \frac{1}{(n-1)^2}, \text{ etc.} \quad . \qquad . \quad (23.21)$$

and, for the z-distribution, about the mean

$$\beta_1 = \frac{\rho^2}{(n-1)^3}(\rho^2 - \tfrac{9}{16}) + \text{terms in } \frac{1}{(n-1)^2}, \text{ etc.} \qquad . \quad (23.22)$$

$$\beta_2 = 3 + \frac{32 - 3\rho^4}{16(n-1)} + \text{terms in } \frac{1}{(n-1)^2}, \text{ etc.} \qquad . \quad (23.23)$$

For $n = 11$, say, β_1 is of the order of 0.001 even if ρ is high, which shows how closely the z-distribution lies to the symmetrical; and $\beta_2 - 3$ is of the order of 0.2, which shows that the distribution has nearly normal kurtosis. In such a case $\bar{z}$ would differ from ζ by 0.05, which is not large, but might be important in some cases.   The standard error of z is, however, $\dfrac{1}{\sqrt{n-3}}$, and the factor $\dfrac{\rho}{2(n-1)}$ may, as a rule, be neglected in comparison.   This is the basis of the statement above that z is normally distributed about mean ζ.

We now give some examples of the use of the z-transformation in testing the significance of an observed r.

Example 23.5.—In Example 11.1, page 215, we found that the correlation between the price indices of animal feeding-stuffs and home-grown oats is 0.68, the sample consisting of 60 members.

This sample is large enough for us to use the standard error.   If we do so we get

$$\sigma_r = \frac{1 - (0.68)^2}{\sqrt{60}} = 0.07 \text{ approximately}$$

The correlation thus is undoubtedly significant.

---

* This z is to be distinguished from the z of Fisher's distribution of **23.24.**

We might, alternatively, use the z test, thus, to answer the question, " Could the observed value have arisen from an uncorrelated universe ? "

On this hypothesis

$$\rho = 0 \quad \text{,and} \quad \zeta = 0$$

We have :

$$z = \tfrac{1}{2} \log_e \frac{1 \cdot 68}{0 \cdot 32}$$

$$= 0 \cdot 829$$

The standard error of z is $\dfrac{1}{\sqrt{57}} = 0 \cdot 13$.

The deviation of z from ζ is more than six times this, and we conclude that our hypothesis was incorrect, *i.e.* that the universe is correlated.

Example 23.6.—Continuing the previous example, could the observed correlation have arisen from a universe in which $\rho = +0 \cdot 8$?

Here

$$\zeta = \tfrac{1}{2} \log_e \frac{1 + \rho}{1 - \rho} = 1 \cdot 099$$

The deviation of z from ζ is, therefore,

$$1 \cdot 099 - 0 \cdot 829 = 0 \cdot 270$$

This is about twice the standard error of z. It might arise, though rarely, as a sampling fluctuation, and we conclude that ρ is likely to be less than $+0 \cdot 8$.

Example 23.7.—In Example 14.1, page 270, we found a partial correlation of $-0 \cdot 73$ (38 unions) between earnings of agricultural labourers and the percentage of the population in receipt of relief, when the ratio of numbers in receipt of outdoor relief to those relieved in the workhouse was constant. Is this significant, and can it have arisen from a universe in which the real correlation is $-0 \cdot 667$?

Here

$$z = \tfrac{1}{2} \log_e \frac{0 \cdot 27}{1 \cdot 73}$$

$$= -0 \cdot 929$$

ζ for an uncorrelated universe $= 0$

$$\zeta, \text{ if } \rho = -0 \cdot 667 = \tfrac{1}{2} \log_e \frac{0 \cdot 333}{1 \cdot 667}$$

$$= -0 \cdot 805$$

There is one secondary subscript in the partial correlation. Hence, the standard error of $z = \dfrac{1}{\sqrt{38 - 1 - 3}} = 0 \cdot 1715$.

If $\zeta = 0$, the deviation is more than five times the standard error and is undoubtedly significant.   If $\rho = -0.667$, the deviation is less than the standard error and hence may very well have arisen from sampling fluctuations

Application of " Student's " Distribution to Correlation Coefficients.

23.39.   The test we have just given is of general application, but it is worth noticing that if $\rho = 0$, the distribution of the correlation coefficient in small samples from a normal universe may be tested by the " Student " distribution.

In fact, the distribution of the correlation coefficient assumes a particularly simple form for such uncorrelated universes, namely,

$$y = y_0(1 - r^2)^{\frac{n-4}{2}} \qquad . \qquad . \qquad . \qquad (23.24)$$

If we put

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n - 2} \qquad . \qquad . \qquad . \qquad (23.25)$$

then it may be shown that t is distributed in the " Student " form with $n - 2$ degrees of freedom, and its significance may be tested accordingly.

Significance of the Correlation Ratio.

23.40.   The distribution of η^2 in samples from an *uncorrelated* normal universe may be derived from Fisher's z-distribution.   Hence we may test whether an observed value of η^2 is significant of the existence of correlation in the parent, assumed normal or approximately so.

When considering the correlation ratio in **13.6** we saw that for the arrays of x's

$$\sigma_x^2 = \sigma_{ax}^2 + \sigma_{mx}^2$$

where

σ_x^2 is the variance of the whole

σ_{ax}^2 is the variance within arrays

σ_{mx}^2 is the variance of array means

If there are p arrays and n_p is the number of members in the pth array, we may write this:

$$S(x - \bar{x})^2 = S(x - \bar{x}_p)^2 + S\{n_p(\bar{x}_p - \bar{x})^2\} \qquad . \qquad . \qquad (23.26)$$

Now let us regard the arrays as classes, and the items of the arrays as class-members.   Equation (23.26) is then an analysis of the sums of squares of the type which we have studied in the analysis of variance. The numbers n_p are not constant in each class, as was k, but this makes no material difference, and we may apply the results of **23.30** to **23.33**.

Using the corrected variances, we may write the analysis in the following tabular form.

TABLE 23.6.

1.	2.	3.	4.	5.
Sums relating to Variation.	Divisor (Degrees of Freedom).	Sums of Squares.	Quotients.	
Between class means . .	$p-1$	$\underset{p=1}{\overset{p=p}{S}}\{n_p(\bar{x}_p-\bar{x})^2\}$	$\dfrac{N\sigma_x^2\eta_{xy}^2}{p-1}$	
Within classes . . .	$N-p$	$\underset{r=1}{\overset{r=N}{S}}(x_r-\bar{x}_p)^2$	$\dfrac{N\sigma_x^2(1-\eta_{xy}^2)}{N-p}$	
Total . .	$N-1$	$\underset{r=1}{\overset{r=N}{S}}(x_r-\bar{x})^2$		

In column 5 we have anticipated results which are easily proved as follows :—

By definition,

$$S(x-\bar{x})^2 = N\sigma_x^2$$
$$S(x-\bar{x}_p)^2 = N\sigma_{ax}^2 = N\sigma_x^2(1-\eta_{xy}^2)$$

Hence,
$$S\{n_p(\bar{x}_p-\bar{x})^2\} = N\sigma_x^2\eta_{xy}^2$$

Dividing the sums of squares by the appropriate number of degrees of freedom, we get the results of column 5.

Now, if the universe is normal and uncorrelated, the two items in column 5 are not significantly different; for they are independent estimates of the variance of x in the universe, all arrays having the same mean and standard deviation.[1] We may test the significance of their difference by the z-distribution. We have :

$$z = \tfrac{1}{2}\log_e \frac{N\sigma_x^2\eta^2}{p-1} \Big/ \frac{N\sigma_x^2(1-\eta^2)}{N-p}$$

$$= \tfrac{1}{2}\log_e \frac{\eta^2}{1-\eta^2} \cdot \frac{N-p}{p-1} \qquad . \qquad . \qquad . \qquad (23.27)$$

$$\left.\begin{array}{l} \nu_1 = p-1 \\ \nu_2 = N-p \end{array}\right\} \qquad . \qquad . \qquad . \qquad (23.28)$$

In equation (23.27) we have omitted the suffix xy in writing η^2. Clearly a similar test may be applied to η_{yx}^2, p in this case referring to the number of y-arrays.

23.41. From the relation (23.27) between z and η^2 it may be shown that the distribution of η^2, corresponding to that of z given by equation (23.15), is

$$y = y_0(\eta^2)^{\frac{p-3}{2}}(1-\eta^2)^{\frac{N-p-2}{2}} \qquad . \qquad . \qquad (23.29)$$

[1] Strictly speaking, this is only approximately true of arrays of finite width. If the ranges defining the arrays are very broad, the test must be used with reserve.

It will be seen that this involves the number p, $i.e.$ depends on the number of arrays into which the data are grouped. This fact is important, and reveals that the use of the standard error $\dfrac{1-\eta^2}{\sqrt{n}}$, given in **21.27**, can be no more than an approximation at the best ; for that formula does not contain p.

23.42. The tables of the significance points of z are designed mainly for small samples. If the data are grouped, as they must be for the calculation of η^2 to be possible, at least one of ν_1, ν_2 is likely to be large. In such cases, however, interpolation will usually give results accurate enough for the purpose in view. But special tables have been prepared by T. L. Woo and appear in " *Tables for Statisticians and Biometricians, Part 2*," to enable closer approximations to be made without arithmetical labour.

23.43. It is interesting to note that, since η^2 is positive, its mean value will not be zero. The mean value (which differs from the square of the mean value of η) is given by

$$\overline{(\eta^2)} = \frac{p-1}{N-1} \qquad . \qquad . \qquad . \qquad (23.30)$$

Example 23.8.—Let us consider the data of Table 11.3 (correlation between stature of father and stature of son), in which $\eta_{xy} = \eta_{yx} = 0.52$. We know that the distribution is approximately normal, a fact which is borne out by the approximate equality of the two correlation ratios, and hence we may apply the foregoing theory with considerable confidence.

We have, for η_{yx}:

$$\nu_1 = p - 1 = 16$$
$$\nu_2 = N - p = 1078 - 17 = 1061$$
$$z = \tfrac{1}{2} \log_e \frac{(0.52)^2}{1-(0.52)^2} \cdot \frac{1061}{16} = 1.60$$

From Appendix Table 6C we see that the 0·1 per cent. significance points are as follows :—

	$\nu_1 = 12$	$\nu_1 = 24$
$\nu_2 = 60$	0·5992	0·4955
$\nu_2 = \infty$	0·5044	0·3786

The observed z is therefore very strongly significant of correlation in the universe.

Test of Linearity of Regression.

23.44. In **13.7** we saw that the regression of y on x was linear if, and only if, $\eta_{yx}^2 - r^2 = 0$. An important question to decide is, therefore, can an observed value of $\eta^2 - r^2$ have arisen from a universe in which the regression is linear, $i.e.$ the true value is zero ?

This question can be decided by the z test in a similar manner to that of **23.40** and **23.41**. We consider the analysis of the sums of squares of deviations from the regression line into two parts : (1) deviations within arrays, and (2) deviations of means of arrays from the regression line. In

this way it may be shown that the linearity may be tested by taking

$$z = \tfrac{1}{2} \log_e \frac{\eta^2 - r^2}{1 - \eta^2} \cdot \frac{N - p}{p - 2} \qquad . \qquad . \qquad . \quad (23.31)$$

$$\left.\begin{array}{l} \nu_1 = p - 2 \\ \nu_2 = N - p \end{array}\right\} \qquad . \qquad . \qquad . \quad (23.32)$$

Example 23.9.—In considering the correlation between old age, pauperism (x) and the proportion of out-relief (y), Yule found (" *Economic Journal,*" vol. 6, 1896, p. 613)

$$N = 235$$
$$r = +0.34$$
$$\eta_{xy} = \quad 0.46$$
$$\eta_{yx} = \quad 0.39$$

for a grouping of 19 x-arrays and 8 y-arrays. Can the regressions be supposed linear ?

For the x-arrays, $N - p = 216$, $p - 2 = 17$

$$\therefore \quad \frac{\eta^2 - r^2}{1 - \eta^2} = \frac{(0.46)^2 - (0.34)^2}{1 - (0.46)^2} = 0.12177$$

$$z = \tfrac{1}{2} \log_e \left(0.12177 \times \frac{216}{17} \right)$$

$$= 0.218$$

The 5 per cent. point for $\nu_1 = 17$, $\nu_2 = \infty$, is about 0.25, and there is thus no reason to suppose from the observed z that the regression is not linear.

For the y-arrays, similarly, $p - 2 = 6$.

$$z = \tfrac{1}{2} \log_e \left(\frac{(0.39)^2 - (0.34)^2}{1 - (0.39)^2} \cdot \frac{227}{6} \right)$$

$$= 0.244$$

This also will be found to lie within the sampling limits, and the test therefore does not reject the linearity of either regression.

Significance of the Multiple Correlation Coefficient.

23.45. The multiple correlation coefficient is in many ways analogous to the correlation ratio, and we may test its significance by a procedure very similar to that used for the significance of the correlation ratio and regressions.

Consider the regression equation with p variates,

$$x_1 = b_2 x_2 + b_3 x_3 + \ldots + b_p x_p$$

the variates being measured from their means.

We may regard the deviations of observed values of x_1 as composed of two parts : (1) deviations from the values of x_1 given by the regression equation, and (2) deviations of the latter from the mean of x_1. The sum of squares can be analysed accordingly.

The sum of squares of deviations of observed values of x_1 from the mean of $x_1 = N\sigma_1{}^2$, by definition, and has $N-1$ degrees of freedom.

The sum of squares of deviations of observed x_1's from the regression values is $N\sigma_{1.2 \ldots p}^2$ which, by the definition of $R_{1(2 \ldots p)}$, is equal to $N\sigma_1{}^2(1 - R_{1(2 : \ldots p)}^2)$. This has $N-p$ degrees of freedom, for $\sigma_1{}^2$ has $N-1$ degrees of freedom, $\sigma_{1.2}^2$ has $N-2$ degrees, and so on. Writing R for $R_{1(2 \ldots p)}$, we may express the analysis in the following tabular form :—

1.	2.	3.	4.	5.
Sums relating to Variation.	Degrees of Freedom.	Sums of Squares.	Quotients.	
Between class means (Regression values from mean.)	$p-1$	$R^2 N\sigma_1{}^2$	$\dfrac{R^2}{p-1} \cdot N\sigma_1{}^2$	
Within classes . (Deviations from regression values.)	$N-p$	$(1-R^2)N\sigma_1{}^2$	$\dfrac{1-R^2}{N-p} \cdot N\sigma_1{}^2$	
Total . .	$N-1$	$N\sigma_1{}^2$		

Now if the universe value of R is zero, the corrected variances of column 5 should not differ significantly; for x_1 and $b_2 x_2 + \ldots + b_p x_p$ are then uncorrelated, and hence deviations of x from the regression values are uncorrelated with, and independent of, deviations of the regression values from the mean, the universe being normal.

Hence we may test the significance of R by putting

$$z = \tfrac{1}{2} \log_e \frac{R^2}{1-R^2} \cdot \frac{N-p}{p-1} \qquad . \qquad . \qquad . \quad (23.33)$$

$$\left. \begin{array}{l} \nu_1 = p-1 \\ \nu_2 = N-p \end{array} \right\} \qquad . \qquad . \qquad . \quad (23.34)$$

It will be seen that equation (23.33) is of the same form as equation (23.27). The distributions of R^2 and η^2 are formally identical, and we have, for instance, corresponding to equation (23.30),

$$(\overline{R^2}) = \frac{p-1}{N-1} \qquad . \qquad . \qquad . \quad (23.35)$$

Example 23.10.—In Example 14.3, page 279, we found $R_{1(23)} = 0.74$. Is this significant ?

We have :

$$p = 3, \qquad N = 38$$
$$\nu_1 = 2, \qquad \nu_2 = 35$$
$$z = \tfrac{1}{2} \log_e \left(\frac{(0.74)^2}{1-(0.74)^2} \cdot \frac{35}{2} \right)$$
$$= 1.53$$

For $\nu_1 = 2$, the 0·1 per cent. significance points are :

$$\nu_2 = 30 \qquad 1\text{·}0859$$
$$\nu_2 = 40 \qquad 1\text{·}0552$$

The observed z is well above these values and hence R is significant.

SUMMARY.

1. As an estimate of the mean of the universe we may take the mean of the sample, whether large or small.

2. If the mean of the universe is known, we may take the mean square deviation about that mean as an estimate of the variance of the universe; *i.e.* the estimate is given by

$$\sigma_s^2 = \frac{1}{n} S(x - m)^2$$

3. If the mean of the universe is not known, a preferable estimate of the universe variance is the " corrected " variance of the sample, given by

$$_c\sigma_s^2 = \frac{1}{n-1} S(x - \bar{x})^2$$

4. This estimate is said to have $n - 1$ degrees of freedom.

5. In samples from a normal universe the parameter t, given by

$$t = \frac{\bar{x} - m}{_c\sigma_s} \sqrt{\nu + 1}$$

where $\nu = n - 1$, is distributed according to the law (due to " Student ")

$$y = \frac{y_0}{\left(1 + \dfrac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}}$$

This distribution may be used to give the probability of getting a value of t between specified limits on random sampling.

6. With two samples, $x_1, \ldots x_{n_1}$ and $x_1', \ldots x_{n_2}'$, from the same normal universe, the parameter t defined by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{_c\sigma_s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

where

$$_c\sigma_s^2 = \frac{1}{n_1 + n_2 - 2} \{S(x - \bar{x}_1)^2 + S(x' - \bar{x}_2)^2\} \qquad \text{and} \qquad \nu = n_1 + n_2 - 2$$

is also distributed according to the above law, with ν degrees of freedom.

7. With two samples, as before, with estimated variances

$$_c\sigma_{s_1}^2 = \frac{1}{n_1 - 1} S(x - \bar{x}_1)^2 \qquad _c\sigma_{s_2}^2 = \frac{1}{n_2 - 1} S(x' - \bar{x}_2)^2$$

the parameter
$$z = \tfrac{1}{2}\log_e \frac{{}_c\sigma_{s_1}^2}{{}_c\sigma_{s_2}^2}$$

is distributed according to the law (due to R. A. Fisher)

$$y = y_0 \frac{e^{\nu_1 z}}{(\nu_1 e^{2z} + \nu_2)^{\frac{\nu_1 + \nu_2}{2}}}$$

where

$$\nu_1 = n_1 - 1, \qquad \nu_2 = n_2 - 1$$

As usual, this distribution may be used to give the probability of getting a value of z between specified limits on random sampling.

8. If the data are arranged in n classes of k members each, the significance of differences between the classes may be tested by comparing $k\sigma_m^2$ with σ_v^2, where σ_m^2 is the variance of class means about the mean of the whole, and σ_v^2 is the average of the variances within classes.

If the sample is small, the comparison may be carried out by applying the z test to the " corrected " variances ${}_c\sigma_m^2$ and ${}_c\sigma_v^2$ with $n-1$ and $n(k-1)$ degrees of freedom respectively, the parent universe being assumed normal.

9. The distribution of the correlation coefficient in samples from a normal bivariate universe is not normal.   However, putting

$$z = \tfrac{1}{2}\log_e \frac{1+r}{1-r}$$

$$\zeta = \tfrac{1}{2}\log_e \frac{1+\rho}{1-\rho}$$

where ρ is the correlation in the universe, it may be shown that z is approximately normally distributed about ζ with standard deviation $\dfrac{1}{\sqrt{n-3}}$, n being the number in the sample.

10. This result remains true of partial correlation coefficients, but in the above formulæ n must be taken to be the number in the sample less the number of secondary subscripts in the coefficient tested.

11. In samples from an uncorrelated normal universe the distribution of r is given by

$$y = y_0(1 - r^2)^{\frac{n-4}{2}}$$

The parameter t, defined by

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n - 2}$$

is distributed in the " Student " form in such cases with $n-2$ degrees of freedom.

12. The significance of η^2 from an uncorrelated normal population may be tested in Fisher's distribution by putting

$$z = \tfrac{1}{2} \log_e \frac{\eta^2}{1 - \eta^2} \cdot \frac{N - p}{p - 1}$$

$$\nu_1 = p - 1, \qquad \nu_2 = N - p$$

where N is the total number in the sample and there are p arrays.

13. The same formulæ give a test for the multiple correlation coefficient R, from a normal universe, if R^2 be substituted for η^2, p being the *total number* of subscripts to R.

14. The linearity of regression in a normal universe, as judged from the value of $\eta^2 - r^2$, may similarly be tested in the z distribution by putting

$$z = \tfrac{1}{2} \log_e \frac{\eta^2 - r^2}{1 - \eta^2} \cdot \frac{N - p}{p - 2}$$

$$\nu_1 = p - 2$$
$$\nu_2 = N - p$$

EXERCISES.

23.1. Find "Student's" t for the following variate values in a sample of 10: -6, -4, -3, -2, -2, 0, 1, 1, 3, 5, taking m to be zero, and find from the tables the probability of getting a value of t as great or greater on random sampling from a normal universe.

23.2. A farmer grows crops on two fields, A and B. On A he puts £1 worth of manure per acre and on B £2 worth. The net returns per acre, exclusive of the cost of manure, on the two fields in five years are :

Year.	Field A, £ per Acre.	Field B, £ per Acre.
1	17	18
2	14	16·5
3	21	24
4	18·5	19
5	22	25

Other things being equal, discuss the question whether it is likely to pay the farmer to continue the more expensive dressing. State clearly the assumptions which you make.

23.3. The heights of six randomly chosen sailors are, in inches: 63, 65, 68, 69, 71 and 72. Those of ten randomly chosen soldiers are: 61, 62, 65, 66, 69, 69, 70, 71, 72 and 73. Discuss the light that these data throw on the suggestion that soldiers are, on the average, taller than sailors.

23.4. In the data of Exercise 23.3, use the z-distribution to discuss whether the samples can have come from universes which are identical so far as height distribution is concerned.

23.5. In three samples of 50 lines each from Shakespeare's "Romeo and Juliet" (an early play), the following numbers of weak endings were observed: 7, 9, 10. In three similar samples from "Cymbeline" (late), the numbers of weak endings were 15, 11, 12. Discuss the suggestion that Shakespeare's prosody, as judged by the number of weak endings, changed with advancing years.

23.6. A random sample of 15 from a normal universe gives a correlation coefficient of -0.5. Is this significant of the existence of correlation in the universe?

23.7. Show that in samples of four from an uncorrelated normal universe all values of the correlation coefficient are equally probable; and that for samples of less than four a zero coefficient is the most improbable.

23.8. What is the probability that a correlation coefficient of $+0.75$ or less can arise in a sample of 30 from a normal universe in which the true correlation is $+0.9$? Compare this with the result given by assuming the sampling distribution normal with standard deviation $\dfrac{1-r^2}{\sqrt{n}}$.

23.9. Test the significance of the partial correlation coefficients of Example 14.1, page 270.

23.10. Test the significance of the two multiple correlation coefficients of Example 14.3, page 279, other than the one tested in Example 23.10.

23.11. Show that in samples of 25 from an uncorrelated normal universe the chance is 1 in 100 that r is greater than about 0·43.

23.12. Referring to Exercise 13.1, test the linearity of the regressions of the distribution of cows in Table 11.4, page 200.

INTERPOLATION AND GRADUATION.

Simple Interpolation.

24.1. If the value of a function of a single variable x, say u_x, has been tabulated for equidistant values of the variable x, $x+h$, $x+2h$, etc., we often require to find the value of the function corresponding to an intermediate value of the variable. Functions in very general use, such as common logarithms, have usually been tabulated with intervals so small that even over a range of several intervals the relation between u_x and x may be assumed to be effectively linear, that is of the form

$$u_x = a_0 + a_1 x \qquad . \qquad . \qquad . \qquad . \quad (24.1)$$

as is shown by the constancy of the differences between successive values of u. For example,

TABLE 24.1.

Number.	Logarithm.	Difference (+).
30597	4·4856788	
		0·0000142
30598	4·4856930	
		0·0000142
30599	4·4857072	
		0·0000142
30600	4·4857214	
		0·0000142
30601	4·4857356	
		0·0000142
30602	4·4857498	

If we then require, say, the value of log 30600·3, it is sufficient to use the familiar process of **simple interpolation** :

$$\begin{array}{ll} \text{log } 30600 & 4\cdot4857214 \\ 0\cdot3 \times 0\cdot0000142 & 43 \\ \hline & 4\cdot4857257 \end{array}$$

The little multiplication sum is, in most tables, already done for us in the margin.

Differences.

24.2. For any function which has been tabulated to sufficiently fine intervals (within certain limitations) simple interpolation can be used in

this way—it is only a question of making the intervals sufficiently small (see below, **24.16**). But many functions have not been tabulated in such detail, successive differences are *not* equal, and consequently simple interpolation cannot give an accurate result. The problem then arises, how are we to interpolate with reasonable precision? And the answer is given *by proceeding to higher orders of differences,* as they are termed; *i.e.* instead of considering only the differences

$$\Delta_0{}^1 = u_1 - u_0$$
$$\Delta_1{}^1 = u_2 - u_1$$
$$\Delta_2{}^1 = u_3 - u_2$$

etc., we also consider the **second differences**

$$\Delta_0{}^2 = \Delta_1{}^1 - \Delta_0{}^1$$
$$\Delta_1{}^2 = \Delta_2{}^1 - \Delta_1{}^1$$
$$\Delta_2{}^2 = \Delta_3{}^1 - \Delta_2{}^1$$

etc., or even the **third differences, fourth differences,** etc.

24.3. To take an actual example, Table 24.2 shows the squares of the first few natural numbers, together with their first and second differences. Following a practice which is convenient for printing and for most purposes of practical work, each difference is printed, not on a line between the two figures to which it relates, as with the logarithms in Table 24.1 above, but on the same line as the upper figure of the two concerned—the line of the figure subtracted; and as the signs of the differences are constant for each column this sign is simply stated at the top.

TABLE 24.2.

Number. x.	Square. u_x.	First Diff. Δ^1 ($+$).	Second Diff. Δ^2 ($+$).	Third Diff. Δ^3.
0	0	1	2	0
1	1	3	2	0
2	4	5	2	0
3	9	7	2	—
4	16	9	—	—
5	25	—	—	—

Here we see that the *first differences*—the only ones with which we have been concerned hitherto—are no longer constant; but they follow a simple rule, in that they are an arithmetic series, a linear function of x. As a result, the *second differences* are constant, actually $+2$, and consequently the third differences vanish.

24.4. The figures on the first line of such a table are called the **leading term** (0) and the **leading differences** ($+1$, $+2$, 0), and it is evident that, given the leading term and the leading differences, the whole table could be built up by successive addition as far as we pleased, without calculating any square directly except for checking. The series of first

differences would be obtained by adding 2 over and over again, starting from the leading difference 1, *i.e.* $1+2=3$, $3+2=5$, etc. The squares would be given then by adding these differences in succession to the leading term 0 : $0+1=1$; $1+3=4$; $4+5=9$, etc.

Differences of a Polynomial.

24.5. From these results we may conclude quite generally that the second differences of *any* polynomial of the second degree,

$$u_x = a_0 + a_1 x + a_2 x^2 \qquad . \qquad . \qquad . \qquad (24.2)$$

are constant and the third differences vanish. For, if we multiply all the squares in Table 24.2 by any factor a_2, we merely multiply all the differences of every order by the same factor ; and the linear part of the function, $a_0 + a_1 x$, cannot contribute to second differences.

Below we give a similar table, Table 24.3, for the *cubes* of the first few natural numbers, and here it will be seen that *third* differences are constant

TABLE 24.3.

Number. x.	Cube. u_x.	First Diff. $\Delta^1(+)$.	Second Diff. $\Delta^2(+)$.	Third Diff. $\Delta^3(+)$.	Fourth Diff. Δ^4.
0	0	1	6	6	0
1	1	7	12	6	0
2	8	19	18	6	—
3	27	37	24	—	—
4	64	61	—	—	—
5	125	—	—	—	—

and fourth differences vanish. By similar reasoning we may conclude that the third differences of *any* polynomial of the third degree,

$$u_x = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \qquad . \qquad . \qquad . \qquad (24.3)$$

are constant and the fourth differences vanish. The student will be quite correct if he draws the general conclusion that for a polynomial of the rth degree,

$$u_x = a_0 + a_1 x + a_2 x^2 + \ . \ . \ . \ + a_r x^r \qquad . \qquad . \qquad (24.4)$$

the rth differences are constant and the $(r+1)$th differences vanish. To prove this it is only necessary to note that each successive differencing lowers the degree of a polynomial by unity, for the difference of any term x^k is

$$(x+1)^k - x^k = kx^{k-1} + \frac{k(k-1)}{1 \ . \ 2} x^{k-2} + \ . \ . \ . \ + 1$$

which is a polynomial of degree $(k-1)$.

Newton's Formula.

24.6. Evidently these results hold out some possibility of generalising our method of interpolation. If, instead of only considering *two* successive values of u_x, say u_0 and u_1, and using the linear relation between u_x and x

that will reproduce these values to give any required intermediate value of u_x, we can use the polynomial of the second degree which will reproduce *three* adjacent values, u_0, u_1, u_2, or that of the third degree which will reproduce *four*, u_0, u_1, u_2, u_3, and evidently we shall be likely to get much more precise results. But to do this we must be able to obtain the required polynomials in terms of the differences. We shall use the notation already introduced, *i.e.*

x.	Function.	First Diffs.	Second Diffs.	Third Diffs.	Fourth Diffs.
0	u_0	Δ_0^1	Δ_0^2	Δ_0^3	Δ_0^4
1	u_1	Δ_1^1	Δ_1^2	Δ_1^3	—
2	u_2	Δ_2^1	Δ_2^2	—	—
3	u_3	Δ_3^1	—	—	—
4	u_4	—	—	—	—

Further, the common interval for the values of x will be taken as unity, as shown; in practical work this is always treated as the unit until the end of the work, just as the class-interval is so treated when calculating the moments of a frequency-distribution.

24.7. Now write down the leading term and leading differences at the head of a table with spacious columns, as below, up to the leading fourth difference, and fill in the rest of the table working back from right to left. In column 5 for third differences we can fill in only the second space, $\Delta_0^3 + \Delta_0^4$. In column 4 for second differences the second term will be $\Delta_0^2 + \Delta_0^3$ (always adding from the line *above* to the right); the third term will be $\Delta_0^2 + 2\Delta_0^3 + \Delta_0^4$. We leave the student to supply the remainder.

1.	2.	3.	4.	5.	6.
x	u_x.	First Diffs.	Second Diffs.	Third Diffs.	Fourth Diffs.
0	$u_0 = u_0$	Δ_0^1	Δ_0^2	Δ_0^3	Δ_0^4
1	$u_1 = u_0 + \Delta_0^1$	$\Delta_0^1 + \Delta_0^2$	$\Delta_0^2 + \Delta_0^3$	$\Delta_0^3 + \Delta_0^4$	—
2	$u_2 = u_0 + 2\Delta_0^1 + \Delta_0^2$	$\Delta_0^1 + 2\Delta_0^2 + \Delta_0^3$	$\Delta_0^2 + 2\Delta_0^3 + \Delta_0^4$	—	—
3	$u_3 = u_0 + 3\Delta_0^1 + 3\Delta_0^2 + \Delta_0^3$	$\Delta_0^1 + 3\Delta_0^2 + 3\Delta_0^3 + \Delta_0^4$	—	—	—
4	$u_4 = u_0 + 4\Delta_0^1 + 6\Delta_0^2 + 4\Delta_0^3 + \Delta_0^4$	—	—	—	—

Now look at the numerical coefficients in the expressions for u_0, u_1, u_2, etc.; they run

$$1$$
$$1 + 1$$
$$1 + 2 + 1$$
$$1 + 3 + 3 + 1$$
$$1 + 4 + 6 + 4 + 1$$

30

These are familiar figures ; they are the terms in the binomial expansions of $(1+1)^0$, $(1+1)^1$, $(1+1)^2$, $(1+1)^3$, etc. We then have, generally,

$$u_x = u_0 + x\Delta_0^1 + \frac{x(x-1)}{1.2}\Delta_0^2 + \frac{x(x-1)(x-2)}{1.2.3}\Delta_0^3 + \ldots \quad (24.5)$$

where the series of differences may be continued so far as is necessary to give a result of the precision desired. This important equation is known as **Newton's Rule** or **Newton's Formula.** It may be repeated that in this form of the equation the unit of x is the interval. There are many other formulæ of interpolation, but we propose to limit ourselves to this and illustrate its uses.

24.8. It will be seen that, if the series on the right of (24.5) is terminated at Δ_0^r, the expression *is* a polynomial of the rth degree in x, though it is not arranged according to powers of x but according to the successive orders of difference, which is more convenient for our present purpose. This polynomial passes through the $r+1$ successive points $(0, u_0)$, $(1, u_1)$, $(2, u_2)$, . . . (r, u_r). In particular, if the series terminates at Δ_0^1, we have simple interpolation and the polynomial reduces to the straight line passing through $(0, u_0)$ and $(1, u_1)$. If it terminates at Δ_0^2, the series represents a parabola of the second degree passing through the three points $(0, u_0)$, $(1, u_1)$, $(2, u_2)$. If it terminates at Δ_0^3, it represents a polynomial of the third degree passing through the four points $(0, u_0)$, $(1, u_1)$, $(2, u_2)$, $(3, u_3)$; and so on. But the student must remember that even though the polynomial reproduces the values of the function at 0, 1, 2 and 3, it does not *necessarily* closely reproduce the function at intermediate values of x. The whole utility of the formula is dependent on the closeness with which the variable can be represented locally by a polynomial of fairly low degree. Most ordinary functions satisfy this condition when tabulated for small intervals, but occasionally the student may find himself in difficulties. We will give some examples in later sections.

We now proceed to some illustrations, and will give a warning at once : *the student must be very careful as to signs.*

Example 24.1.—Given the cubes below, required to find the cube of 32·4.

We give this first as an example in which the interpolation is *exact,* for the third differences are constant, so that we need not proceed further.

Number.	Cube.	Δ^1 (+).	Δ^2 (+).	Δ^3 (+).
31	29791	2977	192	6
32	32768	3169	198	6
33	35937	3367	204	—
34	39304	3571	—	—
35	42875	—	—	—

As interpolation is exact, it does not matter which term we take as u_0. Supposing we take 32. Thus for 32·4, $x=0\cdot4$, and we have :

$$u_{0\cdot4} = u_0 + 0\cdot4\Delta_0{}^1 + \frac{(0\cdot4)(-0\cdot6)}{1\cdot2}\Delta_0{}^2 + \frac{(0\cdot4)(-0\cdot6)(-1\cdot6)}{1\cdot2\cdot3}\Delta_0{}^3$$

$$= 32768 + 0\cdot4(3169) - 0\cdot12(198) + 0\cdot064(6)$$

$$= 32768 + 1267\cdot6 - 23\cdot76 + 0\cdot384$$

$$= 34012\cdot224$$

This may be verified by direct multiplication, or from Barlow's Tables : the student is recommended to carry out a check by taking 31 as u_0.

Example 24.2.—Given the following cube roots, find the cube root of 102·5. The differences have been written, as is frequently done, without the insertion of the decimal point.

Number.	Cube Root.	Δ^1 (+).	Δ^2 (−).	Δ^3 (+).
101	4·6570095	153192	997	14
102	4·6723287	152195	983	—
103	4·6875482	151212	—	—
104	4·7026694	—	.—	—

Here, if we wish to attain the greatest possible precision and include the third difference, we can only take 101 as u_0; x is then 1·5, and

$$u_{1\cdot5} = u_0 + 1\cdot5\Delta_0{}^1 + 0\cdot375\Delta_0{}^2 - 0\cdot0625\Delta_0{}^3$$

$$= 4\cdot6570095 + 0\cdot02297880 - 0\cdot00003739 - 0\cdot00000009$$

$$= 4\cdot67995082$$

Here we have retained an extra place of decimals throughout the arithmetic in order to get the seventh place correct in the final result, and must round this off to 4·6799508. Even so, we cannot avoid the effect of errors in our data, viz. the errors of rounding off, in the seventh place of decimals, the tabulated cube roots : the seventh place in our answer is still liable to an error of ±1 to ±2 for this reason.

It may be noted that, as differences converge so rapidly in this example, simple interpolation would give an error of little more than a unit in the fifth place of decimals.

Example 24.3.—From the table of Ordinates of the Normal Curve (Appendix Table 1) find the value of the ordinate at $x/\sigma = 0\cdot045$.

We give this example partly as a warning to the student to see that his differences are converging so as to be likely to give a good result. The second difference is numerically much larger than the first, viz. 392 against 199; he must then look at the third as well; if this be large also, he may have to go to a high order of differences to get precision. But the third difference is only +18 and the fourth difference smaller still, so third differences will suffice for the highest precision attainable with the five-figure table. Note that the first difference is negative, the

second negative, the third positive, and since the interval is 0·1, $x = 0·45$, *not* 0·045.

In the difference terms we have retained two decimals beyond the five during the work (separated by a comma) :

$$u_{0.45} = u_6 + 0·45\Delta_0{}^1 - 0·12375\Delta_0{}^2 + 0·0639375\Delta_0{}^3$$
$$= 0·39894 - 0·00089,55 + 0·00048,51 + 0·00001,15$$
$$= 0·39854 \text{ rounded off to the fifth place}$$

Interpolating in the seven-figure table, Table II in " *Tables for Statisticians and Biometricians,*" this is found correct to the last place. It may be noted that, if a calculating machine is used, the products given by successive terms can be cumulated on the machine.

Interpolation of Statistical Series.

24.9. So far we have dealt with straightforward interpolation of tabulated mathematical functions. But interpolation may also be employed on statistical series, or series of figures founded on statistics, provided at least that they run tolerably smoothly. No statistical series or series founded on statistics does, however, run absolutely smoothly, like a mathematical function, unless of course it has been deliberately " graduated " to do so. It must be recognised, therefore, in such cases that we are merely using interpolation as a method of *estimating* the truth ; and the truth in all probability would not and could not be given by any process of interpolation.

The following is an illustration of a series based on statistics.

Example 24.4.—In Part II of the Supplement to the 75th Report of the Registrar-General for England and Wales, abridged life-tables were given for a number of counties, etc. The table below shows the expectation of life at ages 25, 35, etc. to 85, based on the mortality of males in Cambridgeshire in 1910–12, *i.e.* the average number of years that individuals would have lived from the given age onwards, if subjected at each age to the mortality mentioned. Required, to interpolate values for the expectation of life at ages 30, 40, etc.

Age.	Expectation of Life (Males).	Δ^1.	Δ^2.	Δ^3.
25	42·21	− 824	+ 20	+ 34
35	33·97	− 804	+ 54	+ 27
45	25·93	− 750	+ 81	+ 76
55	18·43	− 669	+157	− 3
65	11·74	− 512	+154	—
75	6·62	− 358	—	—
85	3·04	—	—	—
Total . . .	—	−3917	+466	+134
Bottom figures less top	−39·17	+ 466	+134	—

Tables of mathematical functions will often give the differences, but in dealing with data of this kind the student will certainly have to form them himself, and should carry out the check shown. Having formed the column of first differences, he should take the total, of course paying attention to signs. In this case the total of first differences is -3917, or inserting the decimal point, $-39 \cdot 17$. This obviously must be equal to the difference between the bottom figure and the top figure in the preceding column, as we see is the case. The following columns must be checked similarly.

The second differences are considerably smaller than the first differences. Third differences are also small, but rather irregular; it will be found, however, that the contributions of the third differences affect only the second place of decimals in the function, so we ought to attain a very fair result.

To get the figures for ages 30 and 40 we have not much choice and must use the known values at ages 25 to 55. On general grounds it seems best to keep the value of x for which we require u_x near the centre of the values used for interpolation. So the expectation at 50 was determined from the values at 35 to 65, that at 60 from the values at 45 to 75, and that at 70 from the values at 55 to 85. The expectation at 80 was determined with the use of the second difference only from the values at 65, 75, 85.

The work is quite straightforward and the results were : 30, 38·09; 40, 29·90; 50, 22·10; 60, 14·94; 70, 8·99; 80, 4·64. The student may find it instructive to draw a chart.

But some qualms were felt as to how far the results could be trusted. A polynomial is not a very good function to represent an empirical function of the present kind which is slowly dropping to zero (see below, **24.12**). It might possibly be more appropriate to take logarithms of the expectations, interpolate between the logarithms and then convert back into numbers. The test was carried out as a control. The following are then the data and the differences :--

Age.	log (Expectation).	Δ^1.	Δ^2.	Δ^3.
25	1·62542	$-0 \cdot 09432$	$-0 \cdot 02298$	$-0 \cdot 00799$
35	1·53110	$-0 \cdot 11730$	$-0 \cdot 03097$	$-0 \cdot 01662$
45	1·41380	$-0 \cdot 14827$	$-0 \cdot 04759$	$-0 \cdot 00536$
55	1·26553	$-0 \cdot 19586$	$-0 \cdot 05295$	$-0 \cdot 03623$
65	1·06967	$-0 \cdot 24881$	$-0 \cdot 08918$	—
75	0·82086	$-0 \cdot 33799$	—	—
85	0·48287	—	—	—
Total	—	$-1 \cdot 14255$	$-0 \cdot 24367$	$-0 \cdot 06620$
Bottom figures less top	$-1 \cdot 14255$	$-0 \cdot 24367$	$-0 \cdot 06620$	—

The work was done exactly as before, except that the expectation at 80 was obtained with three differences from the given values at 55 to 85. The results differed only very slightly from those obtained before, the following table giving a complete comparison :—

Age.	Interpolation.		Difference.
	Direct.	Logarithmic.	
25	42·21	42·21	—
30	38·09	38·07	− 0·02
35	33·97	33·97	—
40	29·90	29·91	+0·01
45	25·93	25·93	—
50	22·10	22·11	+0·01
55	18·43	18·43	—
60	14·94	14·92	− 0·02
65	11·74	11·74	—
70	8·99	9·00	+0·01
75	6·62	6·62	—
80	4·64	4·63	− 0·01
85	3·04	3·04	—

The differences are almost immaterial.

Notes on the Practical Work.

24.10. *Number of Differences to Use.*—Provided differences converge fairly rapidly and continuously, there is little difficulty in coming to a decision. The student knows to how many digits he desires to be accurate, and it is no use his going on to higher orders of difference which affect only places beyond this ; if he wants four-figure accuracy, it is no good his going on to differences which affect only the sixth and seventh places. To enable him to see more quickly the approximate contribution that a difference of any order will give, the following table of the binomial coefficients may be useful :—

TABLE 24.4. -*Table of the Binomial Coefficients in Newton's Formula from* $x = 0$ *to* $x = 2$ *by Intervals of 0·1.*

x	$\dfrac{x(x-1)}{1\cdot 2}$	$\dfrac{x(x-1)(x-2)}{1\cdot 2\cdot 3}$	$\dfrac{x(x-1)(x-2)(x-3)}{1\cdot 2\cdot 3\cdot 4}$
0	0	0	0
0·1	− 0·045	+0·0285	− 0·0206625
0·2	− 0·08	+0·048	− 0·0336
0·3	− 0·105	+0·0595	− 0·0401625
0·4	− 0·12	+0·064	− 0·0416
0·5	− 0·125	+0·0625	− 0·0390625
0·6	− 0·12	+0·056	− 0·0336
0·7	− 0·105	+0·0455	− 0·0261625
0·8	− 0·08	+0·032	− 0·0176
0·9	− 0·045	+0·0165	− 0·0086625
1·0	0	0	0
1·1	+0·055	− 0·0165	+0·0078375
1·2	+0·12	− 0·032	+0·0144
1·3	+0·195	− 0·0455	+0·0193375
1·4	+0·28	− 0·056	+0·0224
1·5	+0·375	− 0·0625	+0·0234375
1·6	+0·48	− 0·064	+0·0224
1·7	+0·595	− 0·0595	+0·0193375
1·8	+0·72	− 0·048	+0·0144
1·9	+0·855	− 0·0285	+0·0078375
2·0	+1	0	0

A word of warning may, however, be desirable. Because the use of the $(r+1)$th difference would not affect the result in the kth figure, it does not *necessarily* follow that this polynomial value will agree with the true value of the function to the kth figure.

If differences do not converge rapidly and continuously, this is in itself evidence that a polynomial of moderately high order does not fit the function well and high precision cannot be expected. The student may occasionally find himself faced by cases more difficult than those of the foregoing illustrations. For example, here are the initial values of P for values of χ^2 proceeding by unity, and degrees of freedom $\nu = 6$ ($n' = 7$), from Table XII in "*Tables for Statisticians, etc., Part I*":

χ^2.	P.	χ^2.	P.
0	1·000000	5	0·543813
1	0·985612	6	0·423190
2	0·919699	7	0·320847
3	0·808847	8	0·238103
4	0·676676	9	0·173578

If we wish to find by interpolation the value at, say, 0·5, apparently we have no choice but to take our u_0 at zero, for the table starts there. If the student begins work accordingly, he will find his differences not behaving at all nicely; the second leading difference is much greater than the first; the third is a good deal less, but the fourth, fifth and sixth much larger than the third, and it is not until the seventh and higher differences that definite convergence seems to be setting in. If he laboriously works step by step, getting successive approximations to the value of P at 0·5 by using one difference, two differences and so on, he will get a series of *very* slowly converging values :

1.	0·992806
2.	0·999247
3.	0·999658
4.	0·998993
5.	0·998445
6.	0·998131
7.	0·997973
8.	0·997899
9.	0·997865

The true value is 0·997839, and he could have obtained this much quicker by direct calculation; even with the nine differences he has got only four-figure accuracy. But he ought not to have expected a good result if he had taken the trouble to look at the run of the differences. The figures give another useful warning. Using three differences, we have a worse result than when using two only. Increasing the number of differences by one step does not *necessarily* increase precision.

Limitation of the number of differences suitable for use, owing to the effect on differences of errors of rounding off, is considered below **(24.14 and 24.15).**

24.11. *Choice of the Set of u's.*—To interpolate, say, at $x = 2\cdot5$, using third differences, one might employ either the u's at 0, 1, 2, 3, or those at 1, 2, 3, 4, or those at 2, 3, 4, 5; one would not go outside these limits or one would have to *extrapolate* for the value at $2\cdot5$, and that would obviously be unsafe. Which set is it best to choose ? Advice cannot be absolutely definite, but it would seem that usually (but not necessarily) values about equidistant from that sought should be equally valuable as guides, and on this principle we should try and keep the value sought so far as possible central to the set of u's employed.

This suggests that *one* reason for our getting so poor a result above was' that we used such a lop-sided set of u's, with the value sought apparently unavoidably near one end. Let us avoid this by a device. Repeat the value of P for $+1$ at -1 on the other side of zero. (It is true that this has no physical meaning, but the function might conceivably run symmetrically on either side of zero, and its graph has clearly high-order contact with a horizontal tangent at zero.) Now take the four values at -1, 0, $+1$, $+2$ and interpolate, using the resulting three differences only :

χ^2.	P.	Δ^1.	Δ^2.	Δ^3.
-1	$0\cdot985612$	$+0\cdot014388$	$-0\cdot028776$	$-0\cdot022749$
0	1	$-0\cdot014388$	$-0\cdot051525$	—
$+1$	$0\cdot985612$	$-0\cdot065913$	—	—
$+2$	$0\cdot919699$	—	—	—

Interpolating for the value of $u_{1\cdot5}$, we have:

$$u_{1\cdot5} = u_0 + 1\cdot5\Delta_0{}^1 + 0\cdot375\Delta_0{}^2 - 0\cdot0625\Delta_0{}^3$$
$$= 0\cdot997825$$

The true value, as stated above, is $0\cdot997839$, and we have got a closer result by this rearrangement, using third differences only, than we did by using nine differences before.

24.12. *Possible Forms of Polynomials.*—The student may also get into difficulties if he does not bear in mind the forms that polynomials can, and cannot, take ; and if he attempts to use this method of interpolation where the polynomial is unlikely to represent the function well even over a moderate range. A polynomial (parabola) of the second order can take only the form (a) in fig. 24.1. A polynomial of the third order can take the form (b), or the form (c) with a wave in the centre. A polynomial of the fourth order can take a form very much resembling (b), but flatter in the centre, or a form like (c), but with three instead of two half-waves in the middle; and so on. A polynomial *cannot* take the form (1) of a curve tangential or asymptotic to the vertical, like the end near zero of an ideal frequency-curve of the distribution-of-wealth type, or (2) of a curve slowly dropping asymptotically to the horizontal, like a logarithmic curve or the tail of the normal curve—and such functions, mathematical or empirical, are very frequent in statistics. In this latter case it would be more probable that the function could be represented by a function of the form

Then taking logs we have :

$$u = \log_e y = a_0 + a_1 x + a_2 x^2 + \ldots$$

that is to say, we come back to the polynomial. Hence, if the function we are dealing with is tailing slowly away to zero, it is probably best to take logarithms and then interpolate on the logarithms. That is why in Example 24.4 we carried out a check in that way. There, as it happened, the direct method did not lead to bad results, but it is quite possible for it to give a completely nonsensical answer. For example, at the extreme end of the χ^2 table for $\nu = 28$ $(n' = 29)$, we are given only the values of P corresponding to the following values of χ^2 : —

χ^2.	P.	Δ^1.	Δ^2.	Δ^3.
40	0·066128	− 0·059661	+ 0·053601	− 0·047929
50	0·006467	− 0·006060	+ 0·005672	—
60	0·000407	− 0·000388	—	—
70	0·000019	—	—	—

Taking differences as shown and interpolating to get an estimate of the value of P for $\chi^2 = 55$, $i.e.$ $u_{1.5}$, we have :

$$u_{1.5} = u_0 + 1·5\Delta_0{}^1 + 0·375\Delta_0{}^2 - 0·0625\Delta_0{}^3$$
$$= - 0·000268$$

But this is nonsense, for P cannot be negative. The polynomial has done its best : it reproduces the values at 40, 50, 60 and 70—but it can only do this by taking a form like (c) of fig. 24.1 (reversed) with a wave in the centre. It has, as a matter of fact, a minimum at $\chi^2 = 56·6$ and a maximum at $\chi^2 = 65·8$, or at $1·66$ and $2·58$ on the scale of u's with 40 as zero and 10 as the unit interval.

If, instead, we take logarithms of the above values of P, interpolate to third differences and then convert back to numbers, as in Example 24.4, we find $0·001699$ for the required value of P—a value which is rational and is probably not far from the truth. For $\chi^2 = 30$, $P = 0·363218$. Even bringing in this much larger value and using logarithmic interpolation with four

FIG. 24.1.

differences, we find $0·001746$ for the value of P at $\chi^2 = 55$. This suggests that at least we may trust the value to two figures as $0·0017$, which would be sufficient for practice ; but the value has not been checked by direct calculation.

Effect of Errors in u on the Differences.

24.13.—The student may notice and be troubled by the fact that, in the Normal Curve Tables in the Appendix, second differences appear to
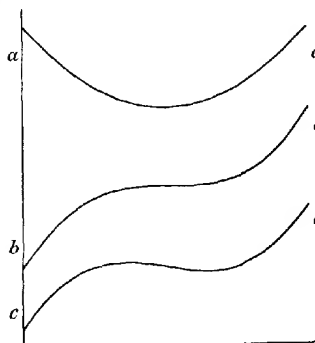
get a little irregular towards the tail of the curve; the phenomenon will become much more evident if he continues the second differences rather further than they have been entered, and still more so in the higher differences if he proceeds to write them out. The irregularities in question are due solely to the errors of rounding off in the last decimal place of the function. Before proceeding to consider the total effect of such a system of errors it may be best to consider the effect of a single error.

24.14. *Effect of an Error in a Single Value of u.*—If $u = v + w$, $\Delta^1 u = \Delta^1 v + \Delta^1 w$, and so on for all orders of differences. Hence, if v represents the true value of u and w represents an error, the differences of the error will simply be superposed on the differences of u, and we may consider the former by themselves. We may then, as below, take the true values of u as zero, and insert an error only at one point, say $+e$.

u.	Δ^1.	Δ^2.	Δ^3.	Δ^4.	Δ^5.	Δ^6.
0	0	0	0	0	0	$+\ e$
0	0	0	0	0	$+\ e$	$-\ 6e$
0	0	0	0	$+\ e$	$-\ 5e$	$+15e$
0	0	0	$+\ e$	$-4e$	$+10e$	$-20e$
0	0	$+\ e$	$-3e$	$+6e$	$-10e$	$+15e$
0	$+e$	$-2e$	$+3e$	$-4e$	$+\ 5e$	$-\ 6e$
$+e$	$-e$	$+\ e$	$-\ e$	$+\ e$	$-\ e$	$+\ e$
0	0	0	0	0	0	0

The resulting differences are written down above, up to those of the sixth order, and it is evident that the numerical coefficients of e in the differences of order r are given by the terms of $(1-1)^r$. The effect of the initial error is therefore very rapidly increased as we proceed to higher and higher orders of difference, especially after the first three differences are past. An error of $+e$ in u can produce an error of $+3e$ or $-3e$ in the third differences, of $6e$ in the fourth differences, of $10e$ in the fifth and of $20e$ in the sixth. The maximum numerical coefficient for order r is derived from that for order $r-1$ by multiplying the latter by 2 if r is even, or by $2r/(r+1)$ if r is odd.

This magnification of the error renders differencing a very useful method of checking the calculated table of a function, and it is often employed for that purpose. The matter is not quite simple, for the effects of errors of rounding off in the last decimal place will be superposed on the effects of any actual mistake, but nevertheless the effects of the mistake are likely to show themselves clearly in, say, third or fourth differences. In the following table of square roots, for example, nothing is obviously wrong, but an error of 2 units in the last place has been introduced into the square root of 15, which should read 3·87298 (or more precisely, 3·8729833). When we proceed to take differences, however, a suspicious irregularity shows itself in the third differences, and in the fourth differences it is clear that something is wrong. Since the position of the " peak " rises half a line at each differencing, the peak $+2$ shows that the mistake is in the root of 15. We can even estimate the magnitude of the error. If the fifth differences may be taken as approximately constant, we ought to get a fair

Number.	Square Root.	Δ^1 (+).	Δ^2 (-).	Δ^3 (+).	Δ^4.
10	3·16228	0·15434	686	83	- 14
11	3·31662	0·14748	603	69	- 12
12	3·46410	0·14145	534	57	- 14
13	3·60555	0·13611	477	43	+ 2
14	3·74166	0·13134	434	45	- 14
15	3·87300	0·12700	389	31	0
16	4	0·12311	358	31	- 6
17	4·12311	0·11953	327	25	—
18	4·24264	0·11626	302	—	—
19	4·35890	0·11324	—	—	—
20	4·47214	—	—	—	—

estimate of the true fourth difference at the peak +2 by adding together that difference and the two on either side of it, the total effect of the error *e* thus averaging out—compare the scheme showing the effect of the single error given above. This average is – 7·6. We then have :

$$6e = +2 - (-7·6)$$
$$e = +1·6$$

This is very near the correct value, which, as will be seen from the true value of the root stated, is $300 - 298·33$ or $1·67$, the unit in the Δ^4 column being the last place of decimals of the function.

24.15. *Effect of a Series of Random Errors in u.*—Suppose these errors to be *a, b, c, d, e,* as below. Writing down their differences, we have the following results :—

Error.	Δ^1.	Δ^2.	Δ^3.	Δ^4.
a	b – a	c – 2b + a	d – 3c + 3b – a	e – 4d + 6c – 4b + a
b	c – b	d – 2c + b	e – 3d + 3c – b	—
c	d – c	e – 2d + c	—	—
d	e – d	—	—	—
e	—	—	—	—

The general result is obvious. In differences of the *r*th order, the resultant error in any one difference is the sum of $r + 1$ of the original errors multiplied in succession by the terms in the binomial expansion of $(1 - 1)^r$, or is of the form

$$e_1 - re_2 + \frac{r(r-1)}{1 \cdot 2} e_3 - \frac{r(r-1)(r-2)}{1 \cdot 2 \cdot 3} e_4 + \; . \; . \; . \qquad . \quad (24.6)$$

If the errors *e* are distributed in a purely random way, so that e_k is uncorrelated with e_{k+s}, and if it may be assumed that the mean error is zero, then the mean error in the difference of the *r*th order will also in a long series tend to zero, and the standard deviation, s_r, of the above quantity (24.6) is given by

$$s_r^2 = F(r)s_0^2 \qquad . \qquad . \qquad . \qquad . \quad (24.7)$$

where s_0 is the s.d. of the original errors e, and $F(r)$ is the sum of the squares of the terms in the binomial expansion of $(1-1)^r$.

$F(r)$ increases very rapidly with r.   The following table gives the value of $F(r)$ and of its square root from $r=1$ to $r=6$:—

r.	$F(r)$.	$\sqrt{F(r)}$.
1	2	1·41
2	6	2·45
3	20	4·47
4	70	8·37
5	252	15·87
6	924	30·40

The standard deviation of errors in the fourth differences is therefore over eight times, and in the sixth differences over thirty times, the s.d. of the errors affecting u.

If the decimal place in u be regarded as following the last figure retained, the errors of rounding off that figure may be regarded as uniformly distributed over a range ± 0.5, and their standard deviation, s_0, is therefore $\sqrt{1/12}$ or 0.288675.   This gives the following figures for the s.d. of errors in the successive orders of difference owing to the errors of rounding off in u :—

Order of Difference.	S.d. of Errors.
1	0·41
2	0·71
3	1·29
4	2·42
5	4·58
6	8·77

The effect of the errors of rounding off evidently increases very rapidly with the order of difference. With a mathematical function for which the true differences rapidly and continuously converge, the effect of the errors will in fact soon, so to speak, "take charge"; the *observed* differences will rapidly and steadily diverge, growing larger with each successive differencing. At the same time two other phenomena will show themselves. Looking back at the scheme showing the effect of the errors a, b, c, d, e, it will be seen that in any one *column* the same error enters into successive differences with sign reversed. Also in any one *line* the same error enters into successive differences with sign reversed. Hence, as the effect of errors of rounding off becomes overwhelmingly great, (1) the differences of the same order tend to alternate in sign, (2) differences of successive orders on the same line tend to alternate in sign. If these phenomena start to show themselves, the student may well suspect he has gone too far in his differencing. It is evidently no use proceeding to an order of differences mainly significant of errors.

These results for the effect on differences of a random series of errors have an application, not only to the effect of errors of rounding off in mathematical tables, but also to the theory of the method of differences in correlation (ref. (331)).

Effect on Differences of Subdividing an Interval.

24.16. We mentioned early in this chapter (24.2) that, in general, it would become possible to use simple interpolation alone on a table of a mathematical function provided intervals were made sufficiently fine, but this was not proved. Let us consider the effect on the differences of subdividing an interval; it will suffice to take the case of halving it, and for brevity let us confine ourselves to the first three differences.

In terms of Newton's formula the values of u at 0, 0·5, 1, 1·5, are

$$\left.\begin{aligned}
u_0 &= u_0 \\
u_{0.5} &= u_0 + 0\cdot5\Delta_0{}^1 - 0\cdot125\Delta_0{}^2 + 0\cdot0625\Delta_0{}^3 \\
u_1 &= u_0 + \Delta_0{}^1 \\
u_{1.5} &= u_0 + 1\cdot5\Delta_0{}^1 + 0\cdot375\Delta_0{}^2 - 0\cdot0625\Delta_0{}^3
\end{aligned}\right\} \qquad . \qquad (24.8)$$

If the student will write down these expressions at the left of a sheet of foolscap placed lengthwise, and take the differences in the ordinary way, he will find that the new leading differences for the subdivided series with intervals of half the original interval are given by

$$\left.\begin{aligned}
\delta_0{}^1 &= 0\cdot5\Delta_0{}^1 - 0\cdot125\Delta_0{}^2 + 0\cdot0625\Delta_0{}^3 \\
\delta_0{}^2 &= 0\cdot25\Delta_0{}^2 - 0\cdot125\Delta_0{}^3 \\
\delta_0{}^3 &= 0\cdot125\Delta_0{}^3
\end{aligned}\right\} \qquad . \qquad (24.9)$$

If the Δ's of the original series converge rapidly, an assumption really implied by the fact that we stopped at the third difference, so that we can regard the successive Δ's as of different orders of magnitude, it will be seen that $\delta_0{}^1$ is of the order of magnitude $0\cdot5\Delta_0{}^1$, $\delta_0{}^2$ is of the order of magnitude $0\cdot25\Delta_0{}^2$, and $\delta_0{}^3$ of the order of magnitude $0\cdot125\Delta_0{}^3$. That is to say, the new differences are not only smaller than the original differences, but converge much more rapidly.

If we had divided the original interval into ten instead of only two parts, we could have found the new leading differences in precisely the same way, and would then have obtained the result that $\delta_0{}^1$ was of the order of magnitude $0\cdot1\Delta_0{}^1$, $\delta_0{}^2$ of the order of magnitude $0\cdot01\Delta_0{}^2$, and so on, the general rule being obvious. Hence it is only necessary to subdivide the interval sufficiently in order to render the differences so rapidly convergent that first differences alone can be used.

In works on the method of differences, tables will usually be found giving for various values of the number of subdivisions the formulæ relating the δ's to the Δ's.

We now turn to some statistical problems.

Breaking up a Group.

24.17. Suppose we are given the numbers living, or the numbers of deaths, in successive ten-year age-groups, we may often desire to estimate the numbers in smaller, *e.g.* five-year, age-groups, or even at single years

of age. The initial difficulty and the method of procedure will best be shown by an illustration.

Example 24.5.—The following are the numbers of deaths in four successive ten-year age-groups. Required to estimate the numbers of deaths at 45–50 and 50–55.

Age-group.	Deaths.
25–	13,229
35–	18,139
45–	24,225
55–	31,496

Now evidently interpolating directly between these figures will not help us. If we interpolated directly between the figure for 35– and the figure for 45– (half-way between), we would only have an estimate of the numbers in the *ten*-year age-group 40–50. We must proceed as follows. Add up the given numbers step by step ; this will give us a new set of figures showing the numbers over 25 but less than 35, over 25 but less than 45, over 25 but less than 55, and over 25 but less than 65. Interpolate in this new series to find the number over 25 but less than 50, and the differences from the numbers next above and below will give the answer desired. The work is as follows :—

1.	2.	3.	4.	5.
Exact Age.	Sum of Deaths from 25 to Age Stated.	Δ^1.	Δ^2.	Δ^3.
25	0	+13,229	+4,910	+1,176
35	13,229	+18,139	+6,086	+1,185
45	31,368	+24,225	+7,271	—
55	55,593	+31,496	—	—
65	87,089	—	—	—

Column 2 gives the numbers from age 25 up to each age stated ; column 3 the first differences, reproducing the numbers in the age-groups ; columns 4 and 5 the second and third differences. Since the two third differences are very nearly equal, working to third differences ought to give us a very fair result. We can accordingly take age 35 as our zero, and age 50 will be 1·5 on the scale with the interval as unit. We have accordingly,

$$u_{1.5} = u_0 + 1 \cdot 5 \Delta_0{}^1 + 0 \cdot 375 \Delta_0{}^2 - 0 \cdot 0625 \Delta_0{}^3$$
$$= 13,229 + 1 \cdot 5(18,139) + 0 \cdot 375(6,086) - 0 \cdot 0625(1,185)$$
$$= 42,645 \cdot 7$$

or **42,646** to the nearest unit. Subtracting 31,368 from 42,646, and

42,646 from 55,593, we then have for our estimates of the numbers of deaths :

$$45\text{--}50 \quad 11,278$$
$$50\text{--}55 \quad 12,947$$

As a matter of fact, the numbers in quinquennial groups were given, and for 45–50, 50–55, were actually 11,404 and 12,821; the error of our estimates accordingly is only of the order of 1 per cent.

Example 24.6.—From the same data, estimate the number of deaths in the year of age 50–51.

The limits of this group on our scale of intervals are, with 35 as origin, 1·5 and 1·6. We have already found the number up to 1·5 in Example 24.5, and it remains only to determine the number up to 1·6, the difference between the two figures then giving the answer sought :

$$
\begin{aligned}
u_{1\cdot6} &= u_0 + 1\cdot6\Delta_0{}^1 + 0\cdot48\Delta_0{}^2 - 0\cdot064\Delta_0{}^3 \\
&= 13,229 + 1\cdot6(18,139) + 0\cdot48(6,086) - 0\cdot064(1,185) \\
&= 45,096\cdot8
\end{aligned}
$$

or 45,097 to the nearest unit. Hence the answer is $45,097 - 42,646$, or 2451.

Simple Formula for Halving a Group.

24.18. The problem of estimating the numbers in the two five-year groups of which a ten-year group is composed occurs so often, that it is worth while deriving a simple second-difference formula for the purpose. Let u's denote numbers in five-year groups, w's numbers in ten-year groups ; and let δ's and Δ's denote the corresponding differences. For second differences we need only consider three consecutive ten-year groups. From Newton's formula we have :

$$
\begin{aligned}
u_0 &= u_0 \\
u_1 &= u_0 + \delta_0{}^1 \\
\hline
w_0 &= 2u_0 + \delta_0{}^1 \\
u_2 &= u_0 + 2\delta_0{}^1 + \delta_0{}^2 \\
u_3 &= u_0 + 3\delta_0{}^1 + 3\delta_0{}^2 \\
\hline
w_1 &= 2u_0 + 5\delta_0{}^1 + 4\delta_0{}^2 \\
u_4 &= u_0 + 4\delta_0{}^1 + 6\delta_0{}^2 \\
u_5 &= u_0 + 5\delta_0{}^1 + 10\delta_0{}^2 \\
\hline
w_2 &= 2u_0 + 9\delta_0{}^1 + 16\delta_0{}^2
\end{aligned}
$$

Now write down these values of the w's and difference :

x.	w_x.	Δ^1.	Δ^2.
0	$2u_0 + \delta_0{}^1$	$4\delta_0{}^1 + 4\delta_0{}^2$	$8\delta_0{}^2$
1	$2u_0 + 5\delta_0{}^1 + 4\delta_0{}^2$	$4\delta_0{}^1 + 12\delta_0{}^2$	
2	$2u_0 + 9\delta_0{}^1 + 16\delta_0{}^2$		

Whence

$$\Delta_0{}^1 = 4(\delta_0{}^1 + \delta_0{}^2)$$
$$\Delta_0{}^2 = 8\delta_0{}^2$$

or

$$\delta_0{}^2 = \tfrac{1}{8}\Delta_0{}^2$$
$$\delta_0{}^1 = \tfrac{1}{4}\Delta_0{}^1 - \tfrac{1}{8}\Delta_0{}^2$$

Hence,

$$u_2 = u_0 + 2\delta_0{}^1 + \delta_0{}^2$$
$$= u_0 + \tfrac{1}{2}\Delta_0{}^1 - \tfrac{1}{8}\Delta_0{}^2$$

$$u_2 - \tfrac{1}{2}w_1 = -\tfrac{1}{8}\Delta_0{}^1 - \tfrac{1}{16}\Delta_0{}^2$$
$$= -\tfrac{1}{16}(2\Delta_0{}^1 + \Delta_0{}^2)$$

It will be convenient for practical work to express this directly in terms of the w's:

$$2\Delta_0{}^1 \quad = 2w_1 - 2w_0$$
$$\Delta_0{}^2 \quad = w_2 - 2w_1 + w_0$$

$$\overline{\quad 2\Delta_0{}^1 + \Delta_0{}^2 = \quad w_2 - w_0 \quad}$$

Whence finally,

$$u_2 = \tfrac{1}{2}\{w_1 + \tfrac{1}{8}(w_0 - w_2)\} \qquad . \qquad . \qquad . \quad (24.10)$$

Thus, taking the figures and problem of Example 24.5 again, we have:

$$w_0 = 18,139$$
$$w_1 = 24,225$$
$$w_2 = 31,496$$

$$\tfrac{1}{8}(w_0 - w_2) = -\ 1,669\cdot6$$
$$w_1 \quad = \quad 24,225$$

$$\overline{\qquad\qquad 22,555\cdot4}$$

and half this gives

$$u_2 = 11,278$$

to the nearest unit, as before. For u_3, of course, we have also, as before, $24,225 - 11,278 = 12,947$. Equation (24.10) is really equivalent to the method of Example 24.5, though in that illustration we used three differences. But the third differences of the numbers "aged over 25 but under x" are equivalent to the second differences of the numbers in the successive age-groups.

Graduation.

24.19. If a graph is drawn showing the numbers of either sex living at each single year of age, as given in any census which provides data in such detail, it will be found anything but smooth, showing the oddest peaks and hollows which repeat themselves, once adult life is reached, at ages showing the same final digits. Thus, in the Census of England and Wales there are conspicuous peaks at the round-numbered ages 30, 40, 50, etc. (last birthday), and hollows or deficiencies at the ages ending with 1 and, less emphatically, at the ages ending with 7. With returns from less

. educated populations, the phenomenon may become almost ludicrous, *e.g.*
in a certain Indian census sample-count :

Age Last Birthday.	Number of Males.
29	927
30	12,294
31	652
32	2,058
33	672
34	892
35	7,723
36	1,437
37	870
38	1,362
39	467
40	10,391
41	460

Now whatever irregularities might occur in the true figures, we may be
quite certain that they should *not* show errors that are simply a function
of the final digit of the age. We would prefer, therefore, to eliminate these
errors. We could do so, somewhat roughly, by drawing a graph as
suggested and sweeping a clean curve through the rather scattered and
irregular points given by the data, subsequently reading off smoothed or
graduated figures from the curve. The graphic process has many points to
recommend it, but is very dependent on personal skill and judgment. It
would be convenient to use a more " mechanical " process that anyone
could apply and be sure of obtaining the same results if he used the same
process. It would be quite possible to fit polynomials to the data by the
methods of Chapter 17, but this would in general entail a great deal of
labour and would not necessarily lead to satisfactory results, *e.g.* with such
highly erratic data as those above. More suitable processes can be
founded on the method of differences, and the general idea of them all is
quite simple, though the details may vary greatly and the practical working
of some of them become rather complex. All methods begin by assuming
that the *totals* of certain age-groups—five-year or ten-year age-groups as
a rule—are reasonably accurate. These totals can then be redistributed
over single years of age by the elementary process of Examples 24.5 and
24.6, or the procedure can be in some way elaborated. We shall illustrate
only the simple process.

Example 24.7.—The English Census of 1911 gives the following numbers
of males in the three age-groups stated. Obtain graduated numbers at
single years of age for the decade 40 to 49.

Age-group.	Number.
30–	2,637,304
40–	2,001,178
50–	1,376,236

As before, we form the sum of these numbers step by step from the top and then take differences.

Exact Age.	Sum of Numbers from 30.	$\Delta^1(+)$.	$\Delta^2(-)$.	$\Delta^3(+)$.
30	0	2,637,304	636,126	11,184
40	2,637,304	2,001,178	624,942	—
50	4,638,482	1,376,236	—	—
60	6,014,718	—	—	—

We now, taking 30 as our zero, require to interpolate at 1·1, 1·2, 1·3, etc. to 1·9. The coefficients of the several differences in the successive applications of Newton's formula are :

Δ^1.	Δ^2.	Δ^3.
+1·1	+0·055	−0·0165
+1·2	+0·12	−0·032
+1·3	+0·195	−0·0455
+1·4	+0·28	−0·056
+1·5	+0·375	−0·0625
+1·6	+0·48	−0·064
+1·7	+0·595	−0·0595
+1·8	+0·72	−0·048
+1·9	+0·855	−0·0285

The results, with the known numbers to age 40 and to age 50 added, are as given in the second column below, and in the fourth column they are differenced to obtain the graduated numbers at each year of age, the total of which must agree with the observed total in the ten-year group.

1.	2.	3.	4.
Exact Age.	Sum of Population from 30 to Age Stated.	Age Last Birthday.	Graduated Number.
40	2,637,304	40	228,559
41	2,865,863	41	222,209
42	3,088,072	42	215,870
43	3,303,942	43	209,542
44	3,513,484	44	203,226
45	3,716,710	45	196,920
46	3,913,630	46	190,626
47	4,104,256	47	184,344
48	4,288,600	48	178,071
49	4,466,671	49	171,811
50	4,638,482		
Total	—	—	2,001,178

Below, these figures are compared with the actual returns at the single years of age and with two other graduations: (1) A graduation given in the Census report and prepared by Mr George King, F.I.A., based on certain quinquennial age-groups. (2) A graduation using analogous methods, but based on ten-year age-groups, made at a later date in the Government Actuary's Department, and reproduced by permission. The methods are described in rather more detail below.

1.	2.	3.	4.	5.
Age Last Birthday.	Census Numbers.	Graduation Above.	King's Graduation, K_1.	Graduation K_2.
40	262,690	228,559	231,070	231,397
41	198,344	222,209	223,721	225,456
42	226,889	215,870	216,556	219,233
43	196,204	209,542	209,314	212,785
44	190,949	203,226	202,143	206,169
45	202,458	196,920	195,193	199,442
46	184,881	190,626	188,610	192,661
47	176,713	184,344	182,577	185,883
48	189,271	178,071	176,994	179,165
49	172,779	171,811	171,589	172,564
Total	2,001,178	2,001,178	1,997,767	2,024,755

If we compare the closeness of fit of the several graduations to the Census returns by adding up the differences, observed number less graduated number, without regard to their sign, and expressing this total as a percentage of the population (2,001,178), it will be found that our graduation gives a percentage deviation of 6·28, King's graduation (K_1) a percentage deviation of 6·09, and the graduation K_2 a percentage deviation of 6·40—figures which do not differ very largely. It will be noticed, however, that both the K graduations give, over the range considered, a small biased error, the total population over the ten years being too small for K_1 and too large for K_2. As regards the deviations of the several graduations from one another, the percentage deviation of our graduation from K_1 is 0·64 and from K_2 1·18, reckoned in each case on the true total population, and the percentage deviation of K_2 from K_1 is 1·35, reckoned on the K_1 total. At some individual ages the differences run up to nearly 2 per cent. This is a warning to the student that while it is true that the use of any one of these methods by different workers must, unlike the use of the graphic method, lead to the same result, yet the choice of *different* methods may lead to results almost, if not quite, as divergent as those obtained by different users of the graphic process. Graduated numbers of hundreds of thousands carried to the last unit suggest a degree of precision much higher than exists.

There is evidently a certain imperfection in the elementary method we have used. If we employed the same method to graduate the numbers at ages 80 to 89, using the numbers in the three ten-year age-groups 20–, 30–, 40–,

there would be a discontinuity at 40, for the two graduated series would be given by arcs of distinct polynomials. The discontinuity might not be conspicuous, but it would be there and would probably be brought out by differencing. To get over this, at least in part, a simple adjustment can be used. Continue the graduated series for 30 to 39 over the next few years of age, say to 42. Also continue our series for 40 to 49 backwards to 37. Over the six years 37 to 42 we then have two graduated values at each age, and these may then be averaged with weights which gradually throw the weight from the earlier series on to the later—say such simple weights as 6 to 1, 5 to 2, 4 to 3, 3 to 4, 2 to 5, 1 to 6. We have also paid no particular attention to the choice of the limits of our ten-year age-group. Of course it might happen that the numbers were only compiled in ten-year groups like 20–, 30–, 40–, etc., and then there would be no choice. But if the figures are given at single years, the choice is at our disposal, and it may be that we have not chosen wisely. Part of the excess at the peak figure is probably drawn from lower ages, and it might have been better to keep the " peak " at the round-number ages well inside the group, e.g. by compiling totals for the decades 35–, 45–, etc., rather than those used.

Mr King, in the Census graduation, used five-year age-groups as his basis, and chose the limits 4–8, 9–13, 14–18, etc., as probably giving the totals nearest the truth. Taking these five-year totals in successive sets of three, he used the precise procedure of our Example 24.6 to determine a graduated figure for the central year of the fifteen—e.g. the three groups covering ages 4–18 would give a graduated number at age 11, the three covering ages 9 to 23 would give a graduated number at age 16, and so on. But here his process broke away. Taking four consecutive graduated numbers five years apart and determined in this way as " pivotal values," he used the method of differences to determine a polynomial of the third order not passing through the four points u_0, u_1, u_2, u_3, but subjected to the four conditions (1) that it should pass through the *two* points u_1 and u_2, (2) that at u_1 and u_2 it should have a common tangent with the corresponding arc determined from the next (overlapping) set of pivotal values. In this way continuity was assured, but equality of observed and graduated totals for the five-year groups was lost. (The process used was a simplification of the process of *osculatory interpolation*, by which two arcs meeting at a point are given not only a common tangent but also a common radius of curvature. It might be called " tangential inter-polation.") The desirability of using five-year groups may be questioned. It is true that ten-year groups are rather large, but the errors that we are trying to eliminate are definitely functions of the ten final digits, and however the limits are chosen there is likely to remain a systematic difference between the adjacent groups of successive pairs if five-year groups are used.

The test of K_2, in which an analogous process was used but based on the ten-year age-groups 5–14, 15–24, etc., was therefore of interest. Over the range of 30–80 years the differences between K_1 and K_2 gave a smoothly running cyclical curve with a tendency towards a period of ten years, as might have been expected.

The simple process given in Example 24.7 is applicable throughout the bulk of life, but not at the two ends of the series, where special tricks of the trade have to be employed. The difficulty of interpolating in a

" tail," where the numbers are slowly approaching zero, has already been pointed out. For graduation these difficulties are increased, and it is often best to drop the method of differences altogether and use some special process, such as assuming a law of decrease or fitting the tail of a frequency-distribution.

Inverse Interpolation.

24.20. By interpolation we determine the value of the function for a given value of the variable. If we are given the value of the function and find the corresponding value of the variable, we are performing **inverse interpolation.** The student has carried out the process, in a form corresponding to simple interpolation, whenever he has determined the number corresponding to a given logarithm by the use of a table of logarithms—not a table of antilogarithms. If we need only take first differences into consideration, the process is, in fact, very simple. From Newton's formula we have

$$u_x = u_0 + x\Delta_0{}^1$$

whence

$$x = \frac{u_x - u_0}{\Delta_0{}^1} \qquad . \qquad . \qquad . \qquad (24.11)$$

where u_0 will naturally be taken as the tabulated value next below u_x.

If we must take second differences also into account, we have

$$u_x = u_0 + x\Delta_0{}^1 + \frac{x(x-1)}{1.2}\Delta_0{}^2$$

which gives the quadratic for x

$$\tfrac{1}{2}\Delta_0{}^2 x^2 + (\Delta_0{}^1 - \tfrac{1}{2}\Delta_0{}^2)x - (u_x - u_0) = 0 \qquad . \qquad . \qquad (24.12)$$

or, solving,

$$x = -\frac{2\Delta_0{}^1 - \Delta_0{}^2}{2\Delta_0{}^2} \pm \sqrt{\frac{2(u_x - u_0)}{\Delta_0{}^2} + \left(\frac{2\Delta_0{}^1 - \Delta_0{}^2}{2\Delta_0{}^2}\right)^2} \qquad . \qquad (24.13)$$

The sign to be taken for the square root will be evident on carrying out the arithmetic.

This is not always a very convenient expression to use, the solution (compare Example 24.8 below) being given as a comparatively small difference between two large quantities. If x_1 is the approximate solution given by first differences, we can replace x in equation (24.12) by $x_1 + h$ and solve for the correction h on the assumption that h^2 may be neglected. This gives

$$h = -\frac{x_1(1 - x_1)\Delta_0{}^2}{2x_1\Delta_0{}^2 + 2\Delta_0{}^1 - \Delta_0{}^2}$$

$$= -\frac{x_1(1 - x_1)\rho}{2 + (2x_1 - 1)\rho} \qquad . \qquad . \qquad . \qquad (24.14)$$

where

$$\rho = \frac{\Delta_0{}^2}{\Delta_0{}^1} . \qquad . \qquad . \qquad . \qquad (24.15)$$

If we may further assume that ρ is small, this reduces to

$$h = \tfrac{1}{2}x_1(1 - x_1)\rho \qquad . \qquad . \qquad . \qquad (24.16)$$

Obtaining a first approximation from first differences, we can use (24.16) to get a second approximation, then insert this second approximation in (24.16) and get a third approximation, and so on until the process of approximation makes no further difference. But note the assumption made that ρ is small.

Example 24.8.—To find from the area-table of the normal curve (Appendix Table 2, p. 532) the approximate value of the quartile deviation, *i.e.* the value of x/σ for which $A = 0.75$.

The data are :

x/σ.	A.	$\Delta_0{}^1$.	$\Delta_0{}^2$.
0·6	0·72575	+ 0·03229	− 0·00219

Hence,

$$u_x - u_0 = 0.02425$$

and the first approximation to x by first differences only is

$$x_1 = + \frac{0.02425}{0.03229} = +0.7510 \text{ interval} \cdot$$
$$= +0.07510$$

or measured from the zero of the scale, the first approximation to the quartile deviation is 0·67510.

Turning now to the quadratic (24.13), the solution is

$$x = 15.2443 - 14.4997$$
$$= 0.7446 \text{ interval}$$
$$= 0.07446$$

the sign of the root having evidently to be taken as negative. Using second differences, then, our approximation to the quartile deviation is

$$0.67446$$

The true value to five places is

$$0.67449$$

so the use of second differences only has left an error in the last digit.

Let us see how the suggested process of approximation would have worked. From (24.16):

$$h = -0.0339114 \times 0.751 \times 0.249$$
$$= -0.00634$$
$$x_1 = \quad 0.751$$
$$\overline{}$$
$$x_2 = \quad 0.74466$$

Now taking x_2 as the second approximation :

$$h = -0.0339114 \times 0.74466 \times 0.25534$$
$$= -0.00645$$
$$x_1 = \quad 0.751$$
$$\overline{}$$
$$x_3 = \quad 0.74455$$

If we repeat the same process again, $x_4 = 0.74455$, which is the same as x_3, so it is no use going further, and 0.67446 is as close as we can get.

If third and higher orders of difference are brought into account, we have an equation of higher degree than the second, which can be solved by Newton's method of approximation, but the student will find more direct methods given in advanced works.

Estimation of the Position of a Maximum.

24.21. In this and the following problem an elementary knowledge of the calculus is assumed ; the student who does not know the calculus may nevertheless find the results useful.

Suppose we are given three equidistant ordinates u_0, u_1, u_2, at 0, 1 and 2. Required to find the position of the maximum of the parabola passing through the tops of the ordinates. We have :

$$u_x = u_0 + x\Delta_0{}^1 + \frac{x(x-1)}{1 \cdot 2}\Delta_0{}^2$$

Differentiating with respect to x and equating to zero, the abscissa of the maximum is given by

$$\Delta_0{}^1 + \tfrac{1}{2}(2x-1)\Delta_0{}^2 = 0$$

or

$$x = 0.5 - \frac{\Delta_0{}^1}{\Delta_0{}^2} \qquad . \qquad . \qquad . \qquad . \qquad (24.17)$$

Very often, perhaps most frequently, our data are not ordinates but rather areas ; *e.g.* if we want to estimate roughly the position of the mode, our data will be the total frequencies in three successive class-intervals—*not* the central ordinates of those intervals. We should then, as in Example 24.5, form the sum of these data step by step and take the *second* differential of the polynomial passing through the resultant points in order to determine the mode. Thus, calling the sum w :

x.	u.	x.	Sum w.
0	u_0	-0.5	0
1	$u_0 + \Delta_0{}^1$	$+0.5$	u_0
2	$u_0 + 2\Delta_0{}^1 + \Delta_0{}^2$	$+1.5$	$2u_0 + \Delta_0{}^1$
		$+2.5$	$3u_0 + 3\Delta_0{}^1 + \Delta_0{}^2$

It must be remembered that the sum w starts at half an interval below zero, as shown. Using δ's to denote the differences of w :

$$\delta_0{}^1 = u_0$$
$$\delta_0{}^2 = \Delta_0{}^1$$
$$\delta_0{}^3 = \Delta_0{}^2$$
$$w_x = w_0 + xu_0 + \frac{x(x-1)}{2}\Delta_0{}^1 + \frac{x(x-1)(x-2)}{6}\Delta_0{}^2$$
$$\frac{d^2w_x}{dx^2} = \Delta_0{}^1 + (x-1)\Delta_0{}^2 = 0$$

or
$$x = 1 - \frac{\Delta_0{}^1}{\Delta_0{}^2}$$

Since x is now measured from $-\frac{1}{2}$, this is the same answer as before.   If we are concerned only with *second* differences of the data, and not with differences of any higher order, it does not matter whether our data are ordinates or areas.

The method must be used with caution ; obviously it cannot give at all a precise result unless the data run smoothly, and if it be used for determining the mode, may easily give an answer appreciably divergent from that obtained by fitting a frequency-curve.   The following illustration will serve as a warning :—

Example 24.9.—The following are the frequencies near the mode in a distribution of barometer heights.   Estimate the position of the mode, (1) from the first three, (2) from the last three.

Height (inches).	Frequency.
29·9	339·5
30·0	382·5
30·1	395·5
30·2	315

Differencing :

Height (inches).	Frequency.	Δ^1.	Δ^2.
29·9	339·5	+43	−30
30·0	382·5	+13	−93·5
30·1	395·5	−80·5	—
30·2	315	—	—

Taking the first three frequencies and their differences :

$$x = 0\cdot5 + \frac{43}{30} = 1\cdot933 \text{ intervals} = 0\cdot193 \text{ inch}$$

$$\therefore \quad \text{Estimated mode} = 30\cdot093$$

Taking the second three frequencies and their differences :

$$x = 0\cdot5 + \frac{13}{93\cdot5} = 0\cdot639 \text{ interval} = 0\cdot064 \text{ inch}$$

$$\therefore \quad \text{Estimated mode} = 30\cdot064$$

Our two answers therefore differ sensibly from each other, and also from the value given by a fitted Pearson curve, viz. 30·039.

Modifying Central Ordinates to Equivalent Areas.

24.22. Supposing we fit a theoretical frequency-curve to an actual distribution, and want to determine the "goodness of fit" by the χ^2 method. We would usually proceed by calculating, from the curve determined, the ordinates at the centre of each class-interval and taking these as the frequencies. But this procedure is not exact, for the central ordinates are not precise measures of the areas. In a class-interval centred exactly on the mode, for example, the central (maximum) ordinate obviously gives too large a value for the area. Required, to obtain some simple formula for modifying the central ordinates so as to give the areas.

We have, by Newton's formula,

$$u_x = u_0 + x\Delta_0{}^1 + \tfrac{1}{2}(x^2 - x)\Delta_0{}^2$$
$$= u_0 + (\Delta_0{}^1 - \tfrac{1}{2}\Delta_0{}^2)x + \tfrac{1}{2}\Delta_0{}^2 x^2$$

Integrate this expression for the interval round u_1, *i.e.* between the limits 0·5 and 1·5, and we will have an expression for the equivalent area, say w_1:

$$w_1 = \int_{0\cdot5}^{1\cdot5} u_x dx = u_0 + \Delta_0{}^1 - \tfrac{1}{2}\Delta_0{}^2 + \tfrac{13}{24}\Delta_0{}^2$$
$$= u_0 + \Delta_0{}^1 + \tfrac{1}{24}\Delta_0{}^2$$
$$\left.\begin{array}{l} w_1 = u_1 + \tfrac{1}{24}\Delta_0{}^2 \\ = \tfrac{1}{24}(u_0 + 22u_1 + u_2) \end{array}\right\} \qquad . \qquad . \qquad (24.18)$$

The first form of the formula is, in general, the more convenient, but the second may be the better if correction is wanted only to a single value of u.

Example 24.10.—Table 24.5 (p. 490) gives in column 2 the calculated ordinates of a Pearson curve at the centres of the class-intervals. In columns 3 and 4 are given the first and second differences, and in column 5 are given the corrections $\Delta_0{}^2/24$, shifted one line down so as to be on the same line as the ordinate to be corrected. Finally, in column 6 we have the sum of the ordinate and the correction, or the area. The totals given at the foot are simply for the purpose of checking; since columns 2 and 3 both begin and end with zero, the sums of both first and second differences must be zero. Since column 5 is derived from column 4 by dividing by 24, its sum should also be zero, but errors of rounding off have made a very small negative excess. All the corrections are very small; they are necessarily greatest where the curvature is greatest.

24.23. A few words in conclusion. The process of interpolation, and still more that of graduation, is almost as much artistic as scientific. No absolute rules can be laid down, judgment must be used, and it is the experienced craftsman who is likely to get the best results with the least labour. If the student turns up his Latin dictionary he will find that *interpolare* means not only " to polish up " (*polire*, to polish)—so that graduation is really the implication of the word—but hence " to corrupt, to falsify." It will do him no harm to bear this etymological meaning in mind, and keep a look-out accordingly.

TABLE 24.5.

1.	2.	3.	4.	5.	6.
Class-interval.	Central Ordinate.	Δ^1.	Δ^2.	Correction.	Area.
—	—	0·00	+ 0·08	—	—
0	0·00	+ 0·08	+ 0·70	+ 0·00	0·00
1	0·08	+ 0·78	+ 3·08	+ 0·03	0·11
2	0·86	+ 3·86	+ 6·91	+ 0·13	0·99
3	4·72	·· 10·77	+ 7·18	+ 0·29	5·01
4	15·49	+ 17·95	− 0·55	+ 0·30	15·79
5	33·44	+ 17·40	− 10·76	− 0·02	33·42
6	50·84	+ 6·64	− 13·70	− 0·45	50·39
7	57·48	− 7·06	− 7·88	− 0·57	56·91
8	50·42	− 14·94	+ 0·06	− 0·33	50·09
9	35·48	− 14·88	+ 4·37	+ 0·00	35·48
10	20·60	− 10·51	+ 4·67	+ 0·18	20·78
11	10·09	− 5·84	+ 3·15	+ 0·19	10·28
12	4·25	− 2·69	+ 1·64	+ 0·13	4·38
13	1·56	− 1·05	+ 0·69	+ 0·07	1·63
14	0·51	− 0·36	+ 0·25	+ 0·03	0·54
15	0·15	− 0·11	+ 0·08	+ 0·01	0·16
16	0·04	− 0·03	+ 0·02	+ 0·00	0·04
17	0·01	− 0·01	+ 0·01	+ 0·00	0·01
18	0·00	0·00	0·00	+ 0·00	0·00
	286·02	+ 57·48	+ 32·89	+ 1·36	286·01
		− 57·48	− 32·89	− 1·37	

SUMMARY.

1. The first, second, third, . . . differences of a function u_x are defined by the equations

$$\Delta_0{}^1 = u_1 - u_0$$
$$\Delta_0{}^2 = \Delta_1{}^1 - \Delta_0{}^1$$
$$\Delta_0{}^3 = \Delta_1{}^2 - \Delta_0{}^2$$
$$\text{etc.}$$

the intervals between successive values of the variable x being equal.

2. By means of Newton's formula,

$$u_x = u_0 + x\Delta_0{}^1 + \frac{x(x-1)}{1 \cdot 2}\Delta_0{}^2 + \frac{x(x-1)(x-2)}{1 \cdot 2 \cdot 3}\Delta_0{}^3 + \; . \; . \; .$$

we can interpolate for the value of u_x.

3. Errors in the values of u become of increasing importance as the order of the differences increases.

4. For inverse interpolation

$$x = \frac{u_x - u_0}{\Delta_0{}^1}$$

for first differences ;

$$x = -\frac{2\Delta_0{}^1 - \Delta_0{}^2}{2\Delta_0{}^2} \pm \sqrt{\frac{2(u_x - u_0)}{\Delta_0{}^2} - \left(\frac{2\Delta_0{}^1 - \Delta_0{}^2}{2\Delta_0{}^2}\right)^2}$$

for second differences.

We can also proceed by successive approximation. If x_1 is the approximate solution by first differences, a closer approximation is $x_1 + h$, where

$$h = \frac{x_1(1 - x_1)\dfrac{\Delta_0{}^2}{\Delta_0{}^1}}{2 + (2x_1 - 1)\dfrac{\Delta_0{}^2}{\Delta_0{}^1}}$$

EXERCISES.

24.1. In the area table of the normal curve, Appendix Table 2, find the value of A for $x/\sigma = 1\cdot 54$, noting the successive approximations up to third differences. Take u_0 at $1\cdot 4$.

24.2. Find as closely as possible the value of P for $\chi^2 = 11\cdot 7$ from the following entries in the χ^2 table ("*Tables for Statisticians*"): $\nu = 17$ ($n' = 18$). Note the successive approximations and the number of places to which your final answer is probably trustworthy.

χ^2.	P.
10	0·903610
11	0·856564
12	0·800136
13	0·736186

24.3. From the following entries in the same table for $\nu = 24$ ($n' = 25$), estimate as closely as you can the value of P for $\chi^2 = 43$. Similarly, estimate the closeness of your approximation.

χ^2.	P.
30	0·184752
40	0·021387
50	0·001416
60	0·000064

24.4. The following (p. 492) were the deaths of males registered in England and Wales during the three years 1930, 1931, 1932, at the ages stated. The figures on the right give the totals of the quinquennial groups which were, on this occasion, held to give the best totals for determining quinquennial "pivotal values." Find graduated numbers for the ages 40 to 44 inclusive.

Age.	Numbers.	Quinquennial Totals.
35	3394	
36	3505	
37	3501	
38	3947	
39	3998	18,345
40	4220	
41	4281	
42	5024	
43	4993	
44	5260	23,778
45	5998	
46	6113	
47	6463	
48	6921	
49	7663	33,158

24.5. Let u_0, u_1, u_2, . . . u_{14} be the numbers in fifteen consecutive years of age, as in Exercise 24.4, and w_0, w_5, w_{10} the totals in the three quinquennial groups. Show that if we want only the graduated figure for u_7 as a "pivotal value," this may be written down at once from the equation

$$u_7 = 0 \cdot 2w_5 - 0 \cdot 008 \Delta^2 w_0$$

(King's formula). Verify by comparison with your answer to Exercise 24.4.

24.6. Generalising the above result, show that if w_0, w_r, w_{2r} are three successive age-groups of r years each, we have for the graduated central value

$$\frac{w_{3r-1}}{2} = \frac{w_r}{r} - \frac{r^2 - 1}{24r^2} \Delta^2 \left(\frac{w_0}{r} \right)$$

and hence if r become indefinitely great, the central ordinate of the middle group of three, with areas w_0, w_1, w_2 and common base c, is given by

$$\frac{w_1}{c} - \frac{1}{24} \Delta^2 \left(\frac{w_0}{c} \right)$$

Verify by finding approximately the central ordinate of the normal curve from the areas between $-0 \cdot 3$ and $-0 \cdot 1$, $-0 \cdot 1$ and $+0 \cdot 1$, $+0 \cdot 1$ and $+0 \cdot 3$ x/σ.

24.7. From the following (abbreviated) entries in the χ^2 table, $\nu = 9$ ($n' = 10$), estimate the value of χ^2 for which $P = 0 \cdot 25$:—

χ^2.	P.
11	0·2757
12	0·2133
13	0·1626

24.8. The next table shows a frequency-distribution of 1000 observations, and also gives the frequencies summed from the top. Estimate (1) the median, (2) the first decile, (3) the ninth decile, (a) as usual by simple interpolation, (b) by bringing second differences also into account.

Interval.	Frequency.	x.	Sum of Frequencies from 0 to x.
0–1	28	1	28
1–2	76	2	104
2–3	114	3	218
3–4	141	4	359
4–5	158	5	517
5–6	142	6	659
6–7	119	7	778
7–8	95	8	873
8–9	63	9	936
9–10	33	10	969
10–11	18	11	987
11–12	8	12	995
12–13	2	13	997
13–14	2	14	999
14–15	—	15	999
15–16	1	16	1000
Total	1000	—	—

24.9. The following are the mean temperatures (Fahrenheit) at Greenwich on three days 30 days apart round the periods of summer maximum and winter minimum. Estimate the approximate dates and values of the maximum and minimum.

Day.	Date.	Temp.	Date.	Temp.
0	15th June	58·8	16th Dec.	40·7
30	15th July	63·4	15th Jan.	38·1
60	14th Aug.	62·5	14th Feb.	39·3

24.10. Taking the value of the central ordinate of the normal curve from Appendix Table 1, estimate the area between the limits $\pm 0 \cdot 1x/\sigma$, and verify your answer from the area table.

REFERENCES.

Since the publication of the first edition of this book the literature of Statistics has grown to such an extent that considerations of space alone would prohibit the inclusion of a complete Bibliography in the present edition. Fortunately, there now appear, from time to time, two reviews of recent advances in Theoretical Statistics, one by J. O. Irwin and others in the *Journal of the Royal Statistical Society*, the other by P. R. Rider in the *Journal of the American Statistical Association*. Both these reviews conclude with lists of references.

In the following lists we have, therefore, attempted to give references to more important Papers published prior to 1932 on subjects mentioned in the text. Some later Papers of special interest, and recent books, have also been included. For subsequent years the student is referred to the reviews by Irwin and Rider mentioned above.

The references are arranged in the following manner : First are given works of general interest on the Theory of Statistics, Probability and related subjects. Then the chapters of the book are dealt with *seriatim*. (This involves certain Papers appearing more than once in the references.) Next come references to certain tables which facilitate calculation, and to tables of functions useful in statistical work. Finally some references are given to Italian statistical literature.

Most of the works cited are to be found in the library of the Royal Statistical Society.

Books on the Theory of Probability.

The student who wishes to proceed to the more advanced theory of statistics will find it necessary to have a good working knowledge of the theory of probability, which lies at the root of most statistical inference from samples. A comprehensive bibliography of the earlier writings on the subject is given in J. M. Keynes' book, No. (8), below.

(1) BACHELIER, L., *Calcul des probabilités*, tome 1; Gauthier-Villars, Paris, 1912.
(2) BACHELIER, L., *Le jeu, la chance, et le hasard*; Flammarion, Paris, 1914.
(3) BERTRAND, J. L. F., *Calcul des probabilités*; Gauthier-Villars, Paris, 1889.
(4) BRUNS, H., *Wahrscheinlichkeitsrechnung und Kollektivmasslehre*; Teubner, Leipzig, 1906.
(5) BURNSIDE, W., *Theory of Probability*; Cambridge University Press, 1928.
(6) HENRY, A., *Calculus and Probability for Actuarial Students*; C. & E. Layton, London, 1922.
(7) JEFFREYS, H., *Scientific Inference*; Cambridge University Press, 1931.
(8) KEYNES, J. M., *A Treatise on Probability*; Macmillan, London, 1921.
(9) LEVY, H., and L. ROTH, *Elements of Probability*; Oxford, The Clarendon Press, 1936.
(10) MISES, R. VON, *Wahrscheinlichkeit, Statistik und Wahrheit*; Springer, Berlin, 1928.
(11) POINCARÉ, H., *Calcul des probabilités*; Gauthier-Villars, Paris, 1896.
(12) VENN, J., *The Logic of Chance : an Essay on the Foundations and Province of the Theory of Probability, with especial reference to its Logical Bearings and its Application to Moral and Social Science and to Statistics*; Macmillan, London, 1888. (Out of print.)

Books on the Theory of Statistics and Combination of Observations.

(13) ANDERSON, O., *Einführung in die mathematische Statistik*; Wien, Julius Springer, 1935.

(14) BROWN, W., and G. H. THOMSON, *The Essentials of Mental Measurement*, 3rd Ed.; Cambridge University Press, 1925.

(15) BRUNT, DAVID, *The Combination of Observations*, 2nd Ed.; Cambridge University Press, 1931.

(16) CZUBER, E., *Wahrscheinlichkeitsrechnung und ihre Anwendung auf Fehlerausgleichung, Statistik und Lebensversicherung*; Teubner, Leipzig, vol. 1, 4th Ed., 1923; vol. 2, 3rd Ed., 1921.

(17) CZUBER, E., *Die statistiche Forschungsmethode*; L. W. Seidel, Wien, 1921.

(18) DARMOIS, G., *Statistique mathématique*; Paris, Libraire Octave Doin, 1928.

(19) ELDERTON, W. PALIN, *Frequency-curves and Correlation*, 2nd Ed.; London, C. & E. Layton, 1927.

(20) EZEKIEL, MORDECAI, *Methods of Correlation Analysis*; John Wiley & Sons, New York; Chapman & Hall, London, 1930. (Full treatment of methods of computation, especially the methods that have been developed by American writers for handling problems with many variables.)

(21) FISHER, ARNE, *The Mathematical Theory of Probabilities and its Application to Frequency-curves and Statistical Methods*, vol. 1; New York (Macmillan), 1915; 2nd Ed., Enlarged, 1922.

(22) FORCHER, HUGO, *Die statistische Methode als selbstandige Wissenschaft*; Leipzig, 1913 (Veit).

(23) JORDAN, CHARLES, *Statistique mathématique*; Gauthier-Villars, Paris, 1927.

(24) KOHN, STANISLAV, *Základy Teorie Statistické Metody (Elements of the Theory of Statistical Method)*, published by the State Statistical Office of the Czechoslovak Republic, Prague, 1929. (A solid work of 483 pp.; detailed bibliographies.)

(25) LEXIS, W., *Abhandlungen zur Theorie der Bevölkerungs under Moralstatistik*; Fischer, Jena, 1903.

(26) MONTESSUS DE BALLORE, R. DE, *Probabilités et Statistiques*; Hermann et Cie, Paris, 1931. (Applications of the binomial series to the fitting of frequency-distributions.)

(27) STEFFENSEN, J. F., *Some Recent Researches in the Theory of Statistics and Actuarial Science*; Cambridge University Press, 1930. (The substance of three lectures delivered in London.)

(28) TSCHUPROW, A. A., *Grundbegriffe und Grundprobleme der Korrelationstheorie*; Teubner, Leipzig, 1925.

(29) WHITTAKER, E. T., and G. ROBINSON, *The Calculus of Observations*; Blackie & Son, London, 2nd Ed., 1932.

Books on Statistical Method.

In certain cases the foregoing references also deal with statistical method. See particularly references (17) and (20).

During recent years interest in statistical method has been evidenced by the issue of a rapidly increasing number of books on the subject. Those in the following list will be found useful as supplementing the present volume :—

(30) DAY, EDMUND E., *Statistical Analysis*; The Macmillan Co., New York, 1925.

(31) FISHER, R. A., *Statistical Methods for Research Workers*; Oliver & Boyd, Edinburgh and London, 6th Ed., 1936.

(32) KELLEY, TRUMAN L., *Statistical Method*; The Macmillan Co., New York, 1923.

(33) MISES, R. VON, *Wahrscheinlichkeitsrechnung und die Anwendung in der Statistik und theoretische Physik*; Deuticke, Wien, 1931.

(34) NICEFORO, A., *La Méthode statistique*; Marcel Giard, Paris, 1925.

(35) PEARSON, E. S., *The Applications of Statistical Methods to Industrial Standardisation and Control*; British Standards Institution, 1936.

(36) RIETZ, H. L., *Mathematical Statistics*; Open Court Publishing Co., Chicago, 1927. (A small work, one of a series intended for those who have some mathematical knowledge but are not specialists. Useful references.)

(37) RIETZ, H. L. (edited by), *Handbook of Mathematical Statistics*; Houghton Mifflin Co., Boston, 1924.

(38) SHEWHART, W. A., *The Economic Control of Quality of the Manufactured Product*; D. van Nostrand Co., New York, 1931; Macmillan, London.

(39) TIPPETT, L. H. C., *The Methods of Statistics*; Williams & Norgate, Ltd., London, 1931. (Useful to the student already possessing some knowledge but who wants an introduction to the methods of R. A. Fisher, analysis of variance, etc. Illustrations mainly biological.)

(40) WESTERGAARD, H., and H. C. NYBØLLE, *Grundzuge der Theorie der Statistik*; Fischer, Jena, 1928.

Vital Statistics.

(41) NEWSHOLME, Sir ARTHUR, *The Elements of Vital Statistics*, Revised Edition; Allen & Unwin, London, 1923.

(42) PEARL, R., *Introduction to Medical Biometry and Statistics*; W. B. Saunders Co., Philadelphia and London, 2nd Ed., Enlarged, 1930.

(43) WHIPPLE, G. C., *Vital Statistics*, 2nd Ed.; Wiley & Sons, New York; Chapman & Hall, London, 1923.

(44) WOODS, HILDA M., and W. T. RUSSELL, *An Introduction to Medical Statistics*; P. S. King & Son, Ltd., London, 1931. (Elementary introduction with reference to statistical methods in general.)

Applications of Statistical Method to Engineering Problems.

This is also a branch on which much work has been done of recent years, but it is one with which we are so wholly unfamiliar that we cannot undertake to give any detailed bibliography. The following books may be found useful, and will give references :—

(45) BECKER, R., H. PLAUT and I. RUNGE, *Anwendungen der mathematischen Statistik*; *auf Probleme der Massenfabrikation*; Julius Springer, Berlin, 1927. (Reprint, 1930.)

(46) FRY, T. C., *Probability and its Engineering Uses*; London, Macmillan & Co.; New York, D. van Nostrand Co., 1928.

(47) KOHLWEILER, EMIL, *Statistik im Dienste der Technik*; R. Oldenbourg, München and Berlin, 1931.

The "Reprints" of the Bell Telephone Laboratories, Incorporated, New York, include a number coming under the present head. Mention may be made in particular of Reprint B-297 (reprinted from the *Journal of the Franklin Institute*, vol. 205, 1928): "Economic Aspects of Engineering Applications of Statistical Methods," by W. A. Shewhart, with a bibliography.

See also the series of Supplements to the *Journal of the Royal Statistical Society* (Industrial and Agricultural Research Section).

Applications of Statistical Method to Agricultural Experiment.

The literature on this subject is enormous. For the general principles of the technique developed in recent years, see—

(48) WISHART, J., and H. G. SANDERS, *Principles and Practice of Field Experimentation*; Empire Cotton Growing Corporation, London, 1935.

Reference may also be made to R. A. Fisher's book, ref. (31) above, and his article on "The Arrangement of Field Experiments" in the *Journal of the Ministry of Agriculture*, vol. 33, 1926–27, p. 503.

See also the series of Supplements to the *Journal of the Royal Statistical Society* (Industrial and Agricultural Research Section).

INTRODUCTION.

The History of the Words " Statistics," " Statistical."

(49) JOHN, V., *Der Name Statistik*; Weiss, Berne, 1883. A translation in *Jour. Roy. Stat. Soc.* for same year.

(50) YULE, G. U., "The Introduction of the Words 'Statistics,' 'Statistical,' into the English Language," *Jour. Roy. Stat. Soc.*, vol. 68, 1905, p. 391.

The History of Statistics in General.

Several works on theory of statistics include short histories, *e.g.* H. Westergaard's *Die Grundzüge der Theorie der Statistik* (Fischer, Jena, 1890), and P. A. Meitzen's *Geschichte, Theorie und Technik der Statistik* (new ed., 1903; American translation by R. P. Falkner, 1891). There is no detailed history in English, but the article "Statistics" in the *Encyclopædia Britannica* (11th ed.) gives a very slight sketch, and the biographical articles in Palgrave's *Dictionary of Political Economy* are useful. Reference may also be made to—

(51) GABAGLIO, ANTONIO, *Teoria generale della statistica*, 2 vols.; Hoepli, Milano, 2nd Ed., 1888. (Vol. 1, *Parte storica*.)

(52) HOTELLING, H., "British Statistics and Statisticians Today," *Jour. Amer. Stat. Assoc.*, vol. 25, 1930, p. 186.

(53) HULL, C. H., *The Economic Writings of Sir William Petty, together with the Observations on the Bills of Mortality more probably by Captain John Graunt*; Cambridge University Press, 2 vols., 1899.

(54) JOHN, V., *Geschichte der Statistik*, 1te Teil, bis auf Quetelet; Enke, Stuttgart, 1884. (All published; the author died in 1900. By far the best history of statistics down to the early years of the nineteenth century.)

(55) KOREN, JOHN, *The History of Statistics, their Progress and Development in Many Countries*; Macmillan Co. (New York), 1918.

(56) MOHL, ROBERT VON, *Geschichte und Litteratur der Staatswissenschaften*, 3 vols.; Enke, Erlangen, 1855–58. (For history of statistics see principally latter half of vol. 3.)

(57) WALKER, HELEN M., *Studies in the History of Statistical Method*; Baltimore, Williams & Wilkins Co., 1929. (Most detailed on recent history: chapters on the Normal Curve, Moments, Percentiles, Correlation, Spearman's Theory of Two Factors for Intelligence, Statistics as a Subject of Instruction in American Universities, and the Origin of certain Technical Terms. Useful bibliographies.)

(57a) WESTERGAARD, H., *Contributions to the History of Statistics*, P. S. King & Sons, 1932.

History of Theory of Statistics.

Somewhat slight information is given in the general works cited. From the purely mathematical side the following are important:—

(58) PEARSON, KARL, "Historical Note on the Origin of the Normal Curve of Errors," *Biometrika*, vol. 14, 1924, p. 402.

(59) PEARSON, KARL, "Notes on the History of Correlation," *Biometrika*, vol. 13, 1920, p. 25.

(60) PEARSON, KARL, "The Contribution of Giovanni Plana to the Normal Bivariate Frequency Surface," *Biometrika*, vol. 20A, 1928, p. 295.

(61) PEARSON, KARL, "James Bernouilli's Theorem," *Biometrika*, vol. 17, 1925, p. 201.

(62) PEARSON, KARL, "Historical Note on the Distributions of Standard Deviations of Samples," *Biometrika*, vol. 23, 1931, p. 416.

(63) TODHUNTER, I., *A History of the Mathematical Theory of Probability from the time of Pascal to that of Laplace*; Macmillan, 1865.

See also Karl Pearson, *The Life, Letters and Labours of Francis Galton*, vol. 2, Chapter 13; Cambridge University Press, 1935; and vol. 3a, Chapter 14.

A classified survey of the statistical work of the late Karl Pearson will be found in the Obituary by G. Udny Yule: "Obituary Notices of Fellows of the Royal Society," No. 5, December 1936.

History of Official Statistics.

(64) BERTILLON, J., *Cours élémentaire de statistique*; Société d'éditions scientifiques, 1895. (Gives an exceedingly useful outline of the history of official statistics in different countries.) See also (55).

CHAPTER 1.   Theory of Attributes—Notation and Terminology.

(65) JEVONS, W. STANLEY, "On a General System of Numerically Definite Reasoning," *Memoirs of the Manchester Lit. and Phil. Soc.*, 1870. Reprinted in *Pure Logic and other Minor Works*; Macmillan, 1890.

(66) YULE, G. U., "On the Association of Attributes in Statistics, etc.," *Phil. Trans. Roy. Soc.*, Series A, vol. 194, 1900, p. 257.

(67) YULE, G. U., "On the Theory of Consistence of Logical Class-frequencies and its Geometrical Representation," *Phil. Trans. Roy. Soc.*, Series A, vol. 197, 1901, p. 91.

(68) YULE, G. U., "Notes on the Theory of Association of Attributes in Statistics," *Biometrika*, vol. 2, 1903, p. 121. (The first three sections of (68) are an abstract of (66) and (67). The remarks made as regards the tabulation of class-frequencies at the end of (66) should be read in connection with the remarks made at the beginning of (67) and in this chapter: *cf.* footnote on p. 94 of (67).)

Material has been cited from, and reference made to the notation used in—

(69) WARNER, F., and Others, "Report on the Scientific Study of the Mental and Physical Conditions of Childhood"; published by the Committee, Parkes Museum, 1895.

(70) WARNER, F., "Mental and Physical Conditions among Fifty Thousand Children, etc.," *Jour. Roy. Stat. Soc.*, vol. 59, 1896, p. 125.

CHAPTER 2.   Consistence of Data.

(71) BOOLE, G., *Laws of Thought*, 1854 (chapter 19, "Of Statistical Conditions").

(72) MORGAN, A. DE, *Formal Logic*, 1847 (chapter 8, "On the Numerically Definite Syllogism").

Refs. (71) and (72), together with (65), are the classical works with respect to the general theory of numerical consistence. The student will find the two above difficult to follow on account of their special notation, and, in the case of Boole's work, the special method employed.

(73) YULE, G. U., "On the Theory of Consistence of Logical Class-frequencies and its Geometrical Representation," *Phil. Trans., Series A*, vol. 197, 1901, p. 91. (Deals at length with the theory of consistence for any number of attributes, using the notation of the present chapters.)

CHAPTER 3.   Association of Attributes.

(74) GREENWOOD, M., and G. U. YULE, "The Statistics of Anti-typhoid and Anti-cholera Inoculations, and the Interpretation of Such Statistics in General," *Proc. Roy. Soc. of Medicine*, vol. 8, 1915, p. 113. (Cited for the discussion of association coefficients in §4, and the conclusion that none of these coefficients is of much value for comparative purposes in interpreting statistics of the type considered.)

(75) LIPPS, G. F., "Die Bestimmung der Abhängigkeit zwischen den Merkmalen eines Gegenstandes," *Berichte d. math.-phys. Klasse d. kgl. sächsischen Gesellschaft d. Wissenschaften*; Leipzig, Feb. 1905. (Deals with the general theory of the dependence between two characters, however classified; the coefficient of association of 3.15 is suggested independently.)

(76) PEARSON, KARL, "On the Correlation of Characters not Quantitatively Measurable," *Phil. Trans. Roy. Soc.*, Series A, vol. 195, 1900, p. 1.

(77) PEARSON, KARL, and DAVID HERON, "On Theories of Association," *Biometrika*, vol. 9, 1913, pp. 159–332. (A reply to criticisms in ref. (80).)

(78) YULE, G. U., "On the Association of Attributes in Statistics," *Phil. Trans. Roy. Soc.*, Series A, vol. 194, 1900, p. 257. (Deals fully with the theory of association: the association coefficient of 3.15 suggested.)

(79) YULE, G. U., "Notes on the Theory of Association of Attributes in Statistics," *Biometrika*, vol. 2, 1903, p. 121. (Contains an abstract of the principal portions of (78) and other matter.)

(80) YULE, G. U., "On the Methods of Measuring the Association between Two Attributes," *Jour. Roy. Stat. Soc.*, vol. 75, 1912, pp. 579–642. (A critical survey of the various coefficients that have been suggested for measuring association and their properties.)

CHAPTER 4. Partial Association.

(81) YULE, G. U., "On the Association of Attributes in Statistics," *Phil. Trans. Roy. Soc.*, Series A, vol. 194, 1900, p. 257. (Deals fully with the theory of partial as well as of total association, with numerous illustrations: a notation suggested for the partial coefficients.)

(82) YULE, G. U., "Notes on the Theory of Association of Attributes in Statistics," *Biometrika*, vol. 2, 1903, p. 121. (*Cf.* especially §§ 4 and 5 on the theory of complete independence, and the fallacies due to mixing of records.)

CHAPTER 5. Manifold Classification.

Contingency.

(83) LIPPS, G. F., "Die Bestimmung der Abhängigkeit zwischen den Merkmalen eines Gegenstandes," *Berichte der math.-phys. Klasse der kgl. sächsischen Gesellschaft der Wissenschaften*; Leipzig, 1905. (A general discussion of the problems of association and contingency.)

(84) PEARSON, KARL, "On the Theory of Contingency and its Relation to Association and Normal Correlation," *Drapers' Company Research Memoirs, Biometric Series I*; Dulau & Co., London, 1904. (The memoir in which the coefficient of contingency is proposed.)

(85) PEARSON, KARL, "On a Coefficient of Class Heterogeneity or Divergence," *Biometrika*, vol. 5, 1906, p. 198. (An application of the contingency coefficient to the measurement of heterogeneity, *e.g.* in different districts of a country, by treating the observed frequencies of some quality $A_1, A_2, \ldots A_x$ in the different districts as rows of a contingency table and working out the coefficient: the same principle is also applicable to the comparison of a single district with the rest of the country.)

(86) PEARSON, KARL, "On the Measurement of the Influence of Broad Categories on Correlation," *Biometrika*, vol. 9, 1913, p. 116.

(87) PEARSON, KARL, "On the General Theory of Multiple Contingency, with Special Reference to Partial Contingency," *Biometrika*, vol. 11, 1915–17, p. 145.

(88) PEARSON, KARL, and J. F. TOCHER, "On Criteria for the Existence of Differential Death-rates," *Biometrika*, vol. 11, 1916, p. 159.

(89) PEARSON, KARL, and E. S. PEARSON, "On Polychoric Coefficients of Correlation," *Biometrika*, vol. 14, 1922, p. 127.

(90) RITCHIE-SCOTT, A., "The Correlation Coefficient of a Polychoric Table," *Biometrika*, vol. 12, 1918, p. 93. (Considers various methods of measuring association with special reference to 4 × 3-fold classifications.)

(91) ROYER, E. B., "A Simple Method for Calculating Mean Square Contingency," *Annals Math. Stats.*, vol. 4, 1933, p. 75.

Isotropy.

(92) YULE, G. U., "On a Property which Holds Good for All Groupings of a Normal Distribution of Frequency for Two Variables, with applications to the Study of Contingency Tables for the Inheritance of Unmeasured Qualities," *Proc. Roy. Soc.*, Series A, vol. 77, 1906, p. 324. (On the property of isotropy and some applications.)

(93) YULE, G. U., "On the Influence of Bias and of Personal Equation in Statistics of Ill-defined Qualities," *Jour. Anthrop. Inst.*, vol. 36, 1906, p. 325. (Includes an investigation as to the influence of bias and of personal equation in creating divergences from isotropy in contingency tables.)

Contingency Tables of Two Rows Only.

(94) PEARSON, KARL, "On a New Method of Determining Correlation between a Measured Character *A* and a Character *B* of which only the Percentage of Cases wherein *B* exceeds (or falls short of) a Given Intensity is recorded for each Grade of *A*," *Biometrika*, vol. 7, 1909, p. 96. (Deals with a measure of dependence for a common type of table, *e.g.* a table showing the numbers of candidates who passed or failed at an examination, for each year of age. The table of such a type stands between the contingency tables for unmeasured characters and the correlation table (chap. 11) for variables. Pearson's method is based on that adopted for the correlation table, and assumes a normal distribution of frequency (chap. 12) for *B*.)

(95) PEARSON, KARL, "On a New Method of Determining Correlation, when one Variable is given by Alternative and the other by Multiple Categories," *Biometrika*, vol. 7, 1910, p. 248. (The similar problem for the case in which the variable is replaced by an unmeasured quality.)

CHAPTER 6. Frequency-Distributions.

(96) PEARSON, KARL, "Skew Variation in Homogeneous Material," *Phil. Trans. Roy. Soc.*, Series A, vol. 186, 1895, pp. 343–414.

(97) PEARSON, KARL, "Cloudiness: Note on a Novel Case of Frequency," *Proc. Roy. Soc.*, vol. 62, 1897, p. 287.

(98) PEARSON, KARL, "Supplement to a Memoir on Skew Variation," *Phil. Trans. Roy. Soc.*, Series A, vol. 197, 1901, pp. 443–459, and Second Supplement, vol. 206, 1916, p. 429.

(99) PARETO, VILFREDO, *Cours d'économie politique*, 2 vols.; Lausanne, 1896–97. See especially tome 2, livre 3, chapter 1, "La courbe des revenus."

The first four memoirs above are mathematical memoirs on the theory of ideal frequency-curves, the first being the fundamental memoir, and the third and fourth supplementary. The elementary student may, however, refer to them with advantage, on account of the large collection of frequency-distributions which is given. Without attempting to follow the mathematics, he may also note that each of our rough empirical types may be divided into several sub-types, the theoretical division into types being made on different grounds.

The fifth work (99) is cited on account of the author's discussion of the distribution of wealth in a community, to which reference was made in **6.22**.

A number of curious distributions will also be found in- -

(100) NICEFORO, ALFREDO, *La misura della vita*; Turin, Fratelli Bocca, 1923.

In connection with the remarks in **6.7** on the grouping of ages, reference may be made to the following in which a different conclusion is drawn as to the best grouping:—

(101) YOUNG, ALLYN A., "A Discussion of Age Statistics," *Census Bulletin 13*, Bureau of the Census, Washington, U.S.A., 1904.

CHAPTER 7. Averages and Other Measures of Location.

General.

(102) FECHNER, G. T., "Ueber den Ausgangswerth der kleinsten Abweichungssumme, dessen Bestimmung, Verwendung und Verallgemeinerung," *Abh. d. kgl. sächsischen Gesellschaft d. Wissenschaften*, vol. 18 (also numbered 11 of the *Abh. d. math.-phys. Klasse*); Leipzig, 1878, p. 1. (The average defined as the origin from which the dispersion, measured in one way or another, is a minimum: geometric mean dealt with incidentally, pp. 13–16.)

(103) FECHNER, G. T., *Kollektivmasslehre*, herausgegeben von G. F. Lipps; Engelmann, Leipzig, 1897. (Posthumously published: deals with frequency-distributions, their forms, averages and measures of dispersion in general: includes much of the matter of (102).)

(104) ZIZEK, FRANZ, *Die statistischen Mittelwerthe*; Duncker und Humblot, Leipzig, 1908: English translation, *Statistical Averages*, translated with additional notes, etc., by W. M. Persons; Holt & Co., New York, 1913. (Non-mathematical, but useful to the economics student for references cited.)

The Geometric Mean.

(105) CRAWFORD, G. E., "An Elementary Proof that the Arithmetic Mean of any number of Positive Quantities is Greater than the Geometric Mean," *Proc. Edin. Math. Soc.*, vol. 18, 1899–1900.

(106) EDGEWORTH, F. Y., "On the Method of ascertaining a Change in the Value of Gold," *Jour. Roy. Stat. Soc.*, vol. 46, 1883, p. 714. (Some criticism of the reasons assigned by Jevons for the use of the geometric mean.)

(107) GALTON, FRANCIS, "The Geometric Mean in Vital and Social Statistics," *Proc. Roy. Soc.*, vol. 29, 1879. p. 365.

(108) JEVONS, W. STANLEY, *A Serious Fall in the Value of Gold ascertained and its Social Effects set forth*; Stanford, London, 1863. Reprinted in *Investigations in Currency and Finance*; Macmillan, London, 1884. (The geometric mean applied to the measurement of price changes.)

(109) JEVONS, W. STANLEY, "On the Variation of Prices and the Value of the Currency since 1782," *Jour. Roy. Stat. Soc.*, vol. 28, 1865. Also reprinted in volume cited above.

(110) KAPTEYN, J. C., *Skew Frequency-curves in Biology and Statistics*; Noordhoff, Groningen, and Wm. Dawson, London, 1903. (Contains, amongst other forms, a generalisation of McAlister's law; see ref. (111).)

(111) McALISTER, DONALD, "The Law of the Geometric Mean," *Proc. Roy. Soc.*, vol. 29, 1879, p. 367. (The law of frequency to which the use of the geometric mean would be appropriate.)

The Mode.

(112) DOODSON, ARTHUR T., "Relation of the Mode, Median and Mean in Frequency Curves," *Biometrika*, vol. 9, 1916–17, p. 429. (Gives a proof of the relation noted in 7.27.)

(113) PEARSON, KARL, "On the Modal Value of an Organ or Character," *Biometrika*, vol. 1, 1902, p. 260. (A warning as to the inadequacy of mere inspection for determining the mode.)

(114) PEARSON, KARL, "Skew Variation in Homogeneous Material," *Phil. Trans. Roy. Soc.*, Series A, vol. 186, 1895, p. 343. (Definition of mode, p. 345.)

(115) YULE, G. U., "Notes on the History of Pauperism in England and Wales, etc.: Supplementary Note on the Determination of the Mode," *Jour. Roy. Stat. Soc.*, vol. 59, 1896, p. 343. (The note deals with elementary methods of approximately determining the mode: the one-third rule and one other.)

Estimates of Population.

(116) WATERS, A. C., "A Method for estimating Mean Populations in the last Inter-censal Period," *Jour. Roy. Stat. Soc.*, vol. 64, 1901, p. 293.

(117) WATERS, A. C., *Estimates of Population : Supplement to Annual Report of the Registrar-General for England and Wales* (Cd. 2618, 1907, p. cxvii).

For the methods formerly used, see the *Reports of the Registrar-General of England and Wales* for 1907, pp. cxxxii–cxxxiv, and for 1910, pp. xi–xii. Estimates are now based on statistics of births, deaths and migrations. *Cf.* SNOW, ref. (300), for a different method based on the symptoms of growth such as numbers of births or of houses.

Index-numbers.

These were incidentally referred to in 7.34. The general theory of index-numbers and the different methods in which they may be formed are not considered in the present work. The student will find copious references to the literature in the following:—

(118) BENNETT, T. L., "The Theory of Measurement of Changes in the Cost of Living," *Jour. Roy. Stat. Soc.*, vol. 83, 1920, p. 455.

(119) BOWLEY, A. L., "The Influence on the Precision of Index-numbers of the Correlation between the Prices of Commodities," *Jour. Roy. Stat. Soc.*, vol. 89, 1926, p. 300.

(120) BOWLEY, A. L., *Prices and Wages in the United Kingdom*, 1914-20; Oxford, 1920 (Clarendon Press).

(121) BOWLEY, A. L., "The Measurement of Changes in Cost of Living," *Jour. Roy. Stat. Soc.*, vol. 82, 1919, p. 343.

(122) EDGEWORTH, F. Y., "Reports of the Committee appointed for the purpose of investigating the best methods of ascertaining and measuring Variations in the Value of the Monetary Standard," *British Association Reports*, 1887 (p. 247), 1888 (p. 181), 1889 (p. 133), and 1890 (p. 485).

(123) EDGEWORTH, F. Y., Article "Index-numbers" in Palgrave's *Dictionary of Political Economy*, vol. 2; Macmillan, 1925.

(124) EDGEWORTH, F. Y., "The Plurality of Index-numbers," *Economic Journal*, vol. 35, 1925, p. 379.

(125) EDGEWORTH, F. Y., "The Element of Probability in Index-numbers," *Jour. Roy. Stat. Soc.*, vol. 88, 1925, p. 557.

(126) FISHER, IRVING, "The Best Form of Index-number," *Quart. Pub. Amer. Stat. Assoc.*, March 1921, p. 533.

(127) FISHER, IRVING, *The Making of Index-numbers*; Houghton Mifflin Co., Boston and New York, 1922. (Useful as a repertory of formulæ, with tests of the results given on certain American data; otherwise, *cf.* reviews in *Economic Journal*, vol. 33, pp. 90 and 246, and *Jour. Roy. Stat. Soc.*, vol. 86, 1923, p. 424, and vol. 87, 1924, p. 89.)

(128) FLUX, A. W., "The Measurement of Price Changes," *Jour. Roy. Stat. Soc.*, vol. 84, 1921, p. 167.

(129) FOUNTAIN, H., "Memorandum on the Construction of Index-numbers of Prices," Board of Trade *Report on Wholesale and Retail Prices in the United Kingdom*, 1903.

(130) GINI, C., "Quelques considérations au sujet de la construction des nombres indices des prix, etc.," *Metron*, vol. 4, 1924, p. 3.

(131) KNIBBS, G. H., "Prices, Price-indexes, and Cost of Living in Australia," *Commonwealth of Australia, Labour and Industrial Branch, Report No. 1*, 1912.

(132) MARCH, L., "Rapport sur les indices de la situation économique," *Bulletin de l'Institut International de Statistique*, t. 21, 1924, pt. 2, p. 3.

(133) MARCH, L., "Les modes de mesure du mouvement général des prix," *Metron*, vol. 1, No. 4, 1921, p. 40.

(134) MARSHALL, A., *Money, Credit and Commerce*, Macmillan, London, 1923.

(135) PERSONS, W. M., "Fisher's Formula for Index-numbers," *Rev. Econ. Statistics*, vol. 3, 1921, p. 103.

(136) WOOD, FRANCES, "The Course of Real Wages in London, 1900-12," *Jour. Roy. Stat. Soc.*, vol. 77, 1913-14, p. 1.

(137) WORKING CLASSES, COST OF LIVING COMMITTEE, 1918, *Report* (Cd. 8980, 1918), H.M. Stationery Office.

For the student of the cost of living in Great Britain the following are useful:—

(138) "Labour Gazette Index-number: Scope and Method of Compilation," *Lab. Gaz.*, March 1920 and Feb. 1921.

(139) "Final Report on the Cost of Living of the Parliamentary Committee of the Trades Union Congress" (The Committee, 32 Eccleston Sq., London, 1921); critical notices of the same in the *Labour Gazette*, Aug. and Sept. 1921; and review by A. L. Bowley, *Econ. Jour.*, Sept. 1921.

CHAPTER 8. Measures of Dispersion.

General.

(140) FECHNER, G. T., "Ueber den Ausgangswerth der kleinsten Abweichungssumme, dessen Bestimmung, Verwendung und Verallgemeinerung," *Abh. d. kgl. sächs. Ges. d. Wissenschaften*, vol. 18 (also numbered vol. 11 of the *Abh. d. math.-phys. Klasse*); Leipzig, 1878, p. 1.

Standard Deviation.

(141) PEARSON, KARL, "Contributions to the Mathematical Theory of Evolution (i. On the Dissection of Asymmetrical Frequency-curves)," *Phil. Trans. Roy. Soc.*, Series A, vol. 185, 1894, p. 71. (Introduction of the term "standard deviation," p. 80.)

Mean Deviation.

(142) LAPLACE, PIERRE SIMON, Marquis de, *Théorie analytique des probabilités*: *2me supplément*, 1818. (Proof that the mean deviation is a minimum when taken about the median.)
(143) TRACHTENBERG, M. I., "A Note on a Property of the Median," *Jour. Roy. Stat. Soc.*, vol. 78, 1915, p. 454. (A very simple proof of the same property.)

Method of Percentiles, including Quartiles, etc.

(144) GALTON, FRANCIS, "Statistics by Intercomparison, with Remarks on the Law of Frequency of Error," *Phil. Mag.*, vol. 49 (4th Series), 1875, pp. 33–46.
(145) GALTON, FRANCIS, *Natural Inheritance*; Macmillan, 1889. (The method of percentiles is used throughout, with the quartile deviation as the measure of dispersion.)

Relative Dispersion.

(146) PEARSON, KARL, "Regression, Heredity and Panmixia," *Phil. Trans. Roy. Soc.*, Series A, vol. 187, 1896, p. 253. (Introduction of "coefficient of variation," pp. 276–277.)
(147) VERSCHAEFFELT, E., "Ueber graduelle Variabilität von pflanzlichen Eigenschaften," *Ber. deutsch. bot. Ges.*, Bd. 12, 1894, pp. 350–355.

Calculation of Mean, Standard Deviation, or of the General Moments of a Grouped Distribution.

We have given a direct method that seems the simplest and best for the elementary student. A process of successive summation that has some advantages can, however, be used instead. The student will find a convenient description with illustrations in—

(148) ELDERTON, W. PALIN, *Frequency-curves and Correlation*; C. & E. Layton, London, 2nd Ed., 1927.

Effect of Grouping Observations.

(149) BATEN, W. D., "Corrections for Moments of a Frequency-distribution in Two Variables," *Ann. Math. Stats.*, vol. 2, 1931, p. 309.
(150) ELDERTON, W. PALIN, "Adjustments for the Moments of J-shaped Curves," *Biometrika*, vol. 25, 1933, p. 179; followed by KARL PEARSON, "Note on Mr Palin Elderton's Corrections to the Moments of J-curves," *ibid.*, p. 180.
(151) MARTIN, E. S., "On the Correction for the Moment Coefficients of Frequency-distributions when the Start of the Frequency is one of the Characteristics to be Determined," *Biometrika*, vol. 26, 1934, p. 12.
(152) PAIRMAN, ELEANOR, and KARL PEARSON, "On Corrections for the Moment Coefficients of Limited Range Frequency-distributions when there are Finite or Infinite Ordinates and any Slopes at the Terminals of the Range," *Biometrika*, vol. 12, 1918–19, p. 231.
(153) PEARSE, G. E., "On Corrections for the Moment Coefficients of Frequency-distributions when there are Infinite Ordinates at One or Both Terminals of the Range," *Biometrika*, vol. 20A, 1928, p. 314.
(154) PEARSON, KARL, and Others (editorial), "On an Elementary Proof of Sheppard's Formulæ for Correcting Raw Moments, and on other allied points," *Biometrika*, vol. 3, 1904, p. 308.
(155) PEARSON, KARL, "On the Influence of 'Broad Categories' on Correlation," *Biometrika*, vol. 9, 1913, pp. 116–139.

(156) Sheppard, W. F., "On the Calculation of the Average Square, Cube, etc., of a large number of Magnitudes," *Jour. Roy. Stat. Soc.*, vol. 60, 1897, p. 698.

(157) Sheppard, W. F., "On the Calculation of the most probable Values of Frequency Constants for Data arranged according to Equidistant Divisions of a Scale," *Proc. Lond. Math. Soc.*, vol. 29, p. 353.

(158) Sheppard, W. F., "The Calculation of Moments of a Frequency-distribution," *Biometrika*, vol. 5, 1907, p. 450.

Coefficient of Variation.

See ref. (146) above, and

(159) Wilson, G. S., and Others, "The Bacteriological Grading of Milk," Special Report 206 of the *Medical Research Council*, 1935.

CHAPTER 9. Moments and Measures of Skewness and Kurtosis.

Moments.

For the introduction of moments and related coefficients and their use in fitting curves to frequency-distributions, see refs. (216), (217) and (218) of Chapter 10.

For methods of calculation of moments, see—

(160) Elderton, W. Palin, *Frequency-curves and Correlation*; C. & E. Layton, London, 2nd Ed., 1927.

For corrections to the moments, see refs. (149)-(158) of Chapter 8.

Skewness.

See refs. (216), (217) and (218) of Chapter 10, and also—

(161) Hotelling, H., and L. M. Solomons, "The Limits of a Measure of Skewness," *Ann. Math. Stats.*, vol. 3, 1932, p. 141.

Seminvariants.

(162) Craig, C. C., "On a Property of the Seminvariants of Thiele," *Ann. Math. Stats.*, vol. 2, 1931, p. 154.

(163) Thiele, T. N., "Theory of Observations" (English version reprinted in *Ann. Math. Stats.*, vol. 2, 1931, p. 165).

See also refs. (416), (424) and (513).

CHAPTER 10. Three Important Theoretical Distributions— the Binomial, the Normal and the Poisson.

(164) Aitken, A. C., "Some Applications of Generating Functions to Normal Frequency," *Quart. Jour. Math.*, vol. 2, 1931, p. 130.

(165) Bernoulli, J., *Ars conjectandi, opus posthumum: Accedit tractatus de seriebus infinitis, et epistola gallicè scripta de ludo pilae reticularis*, 1713. (A German translation in Ostwald's *Klassiker der exakten Wissenschaften*, Nos. 107 and 108.)

For the early classical memoirs on the normal curve or law of error by Laplace, Gauss and others, see Todhunter's *History*, ref. (63).

(166) Camp, B. H., "The Normal Hypothesis," *Jour. Amer. Stat. Assoc.*, vol. 26, March Supplement, 1931, pp. 222-226.

(167) Czuber, E., *Wahrscheinlichkeitsrechnung*; Teubner, Leipzig. (Deduction of Law of Errors.)

(168) Edgeworth, F. Y., Article on the "Law of Error" in the *Encyclopædia Britannica*, 10th Ed., vol. 28, 1902, p. 280.

(169) EDGEWORTH, F. Y., "The Law of Error," *Cambridge Phil. Trans.*, vol. 20, 1904, pp. 36–65, 113–141 (and an Appendix, pp. 1–14, not printed in the *Cambridge Phil. Trans.*, but issued with Reprints).

(170) GALTON, FRANCIS, *Natural Inheritance*; Macmillan & Co., London, 1889. (Mechanical method of forming a binomial or normal distribution, chap. 5, p. 63. For Pearson's generalised machine, see below, ref. (174).)

(171) GUMBEL, E. J., "La distribuzione dei decessi secondo la legge di Gauss," *Giorn. dell' Ist. Ital. degli Att.*, vol. 3, 1932, pp. 311–342.

(172) NIXON, J. W., "An Experimental Test of the Normal Law of Error," *Jour. Roy. Stat. Soc.*, vol. 76, 1913, pp. 702–706.

(173) PEARSON, KARL, "Historical Note on the Origin of the Normal Curve of Errors," *Biometrika*, vol. 14, 1924, p. 402.

(174) PEARSON, KARL, "Skew Variation in Homogeneous Material," *Phil. Trans. Roy. Soc.*, Series A, vol. 186, 1895, p. 343.

For the generalised binomial machine, see §1. The memoir deals with curves derived from the general binomial, and from a somewhat analogous series derived from the case of sampling from limited material. Supplement to the memoir, *ibid.*, vol. 197, 1901, p. 443. Second Supplement, *ibid.*, vol. 216, 1916, p. 429. For a derivation of the same curves from a modified standpoint, ignoring the binomial and analogous distributions, *cf.* ref. (354).

(175) SHEPPARD, W. F., "On the Application of the Theory of Error to Cases of Normal Distribution and Normal Correlation," *Phil. Trans. Roy. Soc.*, Series A, vol. 192, 1898, p. 101. (Includes a geometrical treatment of the normal curve.)

(176) YULE, G. U., "On the Distribution of Deaths with Age when the Causes of Death act cumulatively, and similar Frequency-distributions," *Jour. Roy. Stat. Soc.*, vol. 73, 1910, p. 26. (A binomial distribution with negative index, and the related curve, *i.e.* a special case of one of Pearson's curves, ref. (174).)

Poisson's Distribution.

(177) BORTKIEWICZ, L. VON, *Das Gesetz der kleinen Zahlen*; Teubner, Leipzig, 1898.

(178) BORTKIEWICZ, L. VON, "Ueber die Zeitfolge Zufälliger Ereignisse," *Bull. de l'Institut Int. de Stat.*, tome 20, 2e livre, 1915.

(179) BORTKIEWICZ, L. VON, "Realismus und Formalismus in der mathematischer Statistik," *Allgemein. Stat. Arch.*, vol. 9, 1916, p. 225. (Continues the discussion initiated by the paper of Miss Whitaker, ref. (190).)

(180) GREENWOOD, M., and G. UDNY YULE, "On the Statistical Interpretation of some Bacteriological Methods employed in Water Analysis," *Journal of Hygiene*, vol. 21, 1917, p. 36. (Applies a criterion developed from Poisson's limit to the discrimination of water analyses; numerous arithmetical examples.)

(181) GREENWOOD, M., and G. U. YULE, "An Enquiry into the Nature of Frequency-distributions representative of Multiple Happenings, with particular reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents," *Jour. Roy. Stat. Soc.*, vol. 83, 1920, p. 255.

(182) MORANT, G., "On Random Occurrences in Space and Time when followed by a Closed Interval," *Biometrika*, vol. 13, 1921, p. 309.

(183) NEWBOLD, ETHEL M., "A Contribution to the Study of the Human Factor in the Causation of Accidents," *Industrial Fatigue Research Board*, Report No. 34, 1926.

(184) NEWBOLD, ETHEL M., "Practical Applications of the Statistics of Repeated Events, particularly to Industrial Accidents," *Jour. Roy. Stat. Soc.*, vol. 90, 1927, p. 487.

(185) POISSON, S. D., *Recherches sur la probabilité des jugements, etc.*; Paris, 1837. (Pp. 205–207.)

(186) RUTHERFORD, E., and H. GEIGER, with a note by H. BATEMAN, "The Probability Variations in the distribution of a-particles," *Phil. Mag.*, Series 6, vol. 20, 1910, p. 698. (The frequency of particles emitted during a small interval of time follows the law of small chances: the law deduced by Bateman in ignorance of previous work.)

(187) SOPER, H. E., "Tables of Poisson's Exponential Binomial Limit," *Biometrika*, vol. 10, 1914, pp. 25–35.

(188) "STUDENT," "On the Error of Counting with a Hæmacytometer," *Biometrika*, vol. 5, 1907, p. 351.

(189) "Student," "An Explanation of Deviations from Poisson's Law in Practice," *Biometrika*, vol. 10, 1919, p. 211.
(190) Whitaker, Lucy, "On Poisson's Law of Small Numbers," *Biometrika*, vol. 10, 1914, pp. 36-71.

Frequency-distributions in General.

(191) Baten, W. D., "Frequency Laws for the Sum of *n* Variables which are subject to Given Frequency Laws," *Metron*, vol. 10, part 3, 1933, p. 75.
(192) Camp, B. H., "Probability Integrals for the Point Binomial," *Biometrika*, vol. 16, 1924, p. 163.
(193) Camp, B. H., "Probability Integrals for a Hypergeometrical Series," *Biometrika*, vol. 17, 1925, p. 61.
(194) Charlier, C. V. L., Numerous papers issued from the Astronomical Department of Lund, 1906-12, especially "Contributions to the Mathematical Theory of Statistics" (1912).
(195) Charlier, C. V. L., "Researches into the Theory of Probability" (*Communications from the Astronomical Observatory*, Lund); Lund, 1906.
(196) Charlier, C. V. L., "A New Form of the Frequency Function," *Meddelande, Lunds Astronomiska Observatorium*, 1928.
(197) Cramér, H., "On some Classes of Series used in Mathematical Statistics," *Den sjette Skandinaviske Matematikercongres*, Copenhagen, 1928.
(198) Cramér, H., "On the Composition of Elementary Errors," *Skandinavisk Aktuarietidskrift*, 1928.
(199) Cunningham, E., "The ω-Functions, a Class of Normal Functions occurring in Statistics," *Proc. Roy. Soc.*, Series A, vol. 81, 1908, p. 310.
(200) Dodd, E. L., "The Frequency Laws of a Function of Variables with given Frequency Laws," *Annals of Mathematics*, vol. 27, 1925, p. 12.
(201) Dodd, E. L., "The Frequency Law of a Function of One Variable," *Bull. Amer. Math. Soc.*, vol. 31, 1925.
(202) Dodd, E. L., "On Ordinary Plane and Skew Curves," *Bulletin of the Univ. of Texas*, No. 222, 1912.
(203) Dodd, E. L., "Classification of Sizes and Measures by Frequency Functions," *Jour. Amer. Stat. Assoc.*, vol. 26, 1931, p. 277. (A survey: useful references.)
(204) Edgeworth, F. Y., "On the Mathematical Representation of Statistical Data," *Jour. Roy. Stat. Soc.*, vol. 79, 1916, p. 456; vol. 80, 1917, pp. 65, 266, 411; vol. 81, 1918, p. 322.
(205) Edgeworth, F. Y., "On the Representation of Statistics by Mathematical Formulæ," *Jour. Roy. Stat. Soc.*, vol. 61, 1898, p. 670; vol. 62, 1899, p. 125; vol. 63, 1900, p. 72.
(206) Edgeworth, F. Y., Article on the "Law of Error" in the *Encyclopædia Britannica*, 10th Ed., vol. 28, 1902, p. 280.
(207) Edgeworth, F. Y., "The Law of Error," *Cambridge Phil. Trans.*, vol. 20, 1904, pp. 36-65, 113-141 (and an Appendix, pp. 1-14, not printed in the *Cambridge Phil. Trans.*, but issued with Reprints).
(208) Edgeworth, F. Y., "The Generalised Law of Error, or Law of Great Numbers," *Jour. Roy. Stat. Soc.*, vol. 69, 1906, p. 497.
(209) Edgeworth, F. Y., "On the Representation of Statistical Frequency by a Curve," *Jour. Roy. Stat. Soc.*, vol. 70, 1907, p. 102.
(210) Edgeworth, F. Y., "Untried Methods of Representing Frequency," *Jour. Roy. Stat. Soc.*, vol. 87, 1924, p. 571.
(211) Edgeworth, F. Y., "Mr Rhodes's Curve and the Method of Adjustment," *Jour. Roy. Stat. Soc.*, vol. 89, 1926, p. 129. (See ref. (221).)
(212) Frisch, R., "On the Use of Difference Equations in the Study of Frequency-distributions," *Metron*, vol. 10, 1933, part 3, p. 35.
(213) Geary, R. C., "The Frequency-distribution of the Quotient of Two Normal Variables," *Jour. Roy. Stat. Soc.*, vol. 93, 1930, p. 442.
(214) Kapteyn, J. C., *Skew Frequency-curves in Biology and Statistics*; Noordhoff, Groningen; Wm. Dawson & Sons, London, 1903.
(215) Nixon, J. W., "An Experimental Test of the Normal Law of Error," *Jour. Roy. Stat. Soc.*, vol. 76, 1913, pp. 702-706.
(216) Pearson, Karl, "Skew Variation in Homogeneous Material," *Phil. Trans. Roy. Soc.*, Series A, vol. 186, 1895, p. 343, and Supplement, vol. 197, 1901, p. 443.

(217) PEARSON, KARL, "Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson: A Rejoinder," *Biometrika*, vol. 4, 1905, p. 169.

(218) PEARSON, KARL, "Second Supplement to a Memoir on Skew Variation," *Phil. Trans. Roy. Soc.*, Series A, vol. 216, 1916, p. 429.

(219) PEARSON, KARL, "Historical Note on the Origin of the Normal Curve of Errors," *Biometrika*, vol. 14, 1924, p. 402.

(220) PEROZZO, LUIGI, "Nuove Applicazioni del Calcolo delle Probabilità allo Studio dei Fenomeni Statistici e Distribuzione dei Matrimoni secondo l'Età degli Sposi," *Mem. della Classe di Scienze morali, etc., Reale Accad. dei Lincei*, vol. 10, Series 3, 1882.

(221) RHODES, E. C., "On the Generalised Law of Error," *Jour. Roy. Stat. Soc.*, vol. 88, 1925, p. 576.

(222) RIETZ, H. L., "On certain Properties of Frequency-distributions obtained by a Linear Fractional Transformation of the Variates of a given Distribution," *Ann. Math. Stats.*, vol. 2, 1931, p. 38.

(223) ROMANOVSKY, V., "Generalisation of some Types of the Frequency-curves of Professor Pearson," *Biometrika*, vol. 16, 1924, p. 106.

(224) SOPER, H. E., *Frequency Arrays*; Cambridge University Press, 1922.

  (The above are concerned with the general theory of frequency systems; the following deal with the forms which are suitable for the representation of particular classes of data, *e.g.* statistics of epidemic diseases, statistics of accidents, etc.)

(225) BROWNLEE, J. "The Mathematical Theory of Random Migration and Epidemic Distribution," *Proc. Roy. Soc. Edin.*, vol. 31, 1910–11, p. 262.

(226) BROWNLEE, J., "Certain Aspects of the Theory of Epidemiology in Special Reference to Plague," *Proc. Roy. Soc. Medicine, Sect. Epidemiology and State Medicine*, vol. 10D, 1918, p. 85. (The appendix to this paper summarises the author's results and those of Sir Ronald Ross; *vide infra*.)

(227) GREENWOOD, M., and G. U. YULE, "An Enquiry into the Nature of Frequency-distributions Representative of Multiple Happenings, with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents," *Jour. Roy. Stat. Soc.*, vol. 83, 1920, p. 255.

(228) KNIBBS, G. H., "The Mathematical Theory of Population," Appendix A to vol. 1 of *Census of the Commonwealth of Australia*. (Contains a full discussion of the application of various frequency systems to vital statistics.)

(229) MOIR, H., "Mortality Graphs," *Trans. Actuarial Soc. America*, vol. 18, 1917, p. 311. (Numerous graphs of mortality rates in different classes and periods.)

(230) ROSS, Sir RONALD, "An Application of the Theory of Probabilities to the Study of *a priori* Pathometry," *Proc. Roy. Soc.*, A, vol. 92, 1916, p. 204.

(231) ROSS, Sir RONALD, and HILDA P. HUDSON, "An Application of the Theory of Probabilities to the Study of *a priori* Pathometry," Pts. 2 and 3, *Proc. Roy. Soc.*, A, vol. 93, 1917, pp. 212 and 225.

The Resolution of a Distribution compounded of Two Normal Curves into its Components.

(232) EDGEWORTH, F. Y., "On the Representation of Statistics by Mathematical Formulæ," Pt. 2, *Jour. Roy. Stat. Soc.*, vol. 62, 1899, p. 125.

(233) HELGUERO, FERNANDO DE, "Per la risoluzione delle curve dimorfiche," *Biometrika*, vol. 4, 1905, p. 230. Also memoir under the same title in the *Transactions of the Accademia Reale dei Lincei, Rome*, vol. 6, 1906. (The first is a short note, the second the full memoir.)

  See also the memoir by Charlier, cited in (195), section 6 of that memoir dealing with the problem of dissection.

(234) PEARSON, KARL, "Contributions to the Mathematical Theory of Evolution" (on the dissection of asymmetrical frequency-curves), *Phil. Trans. Roy. Soc.*, Series A, vol. 185, 1894, p. 71.

(235) PEARSON, KARL, "On some Applications of the Theory of Chance to Racial Differentiation," *Phil. Mag.*, 6th Series, vol. 1, 1901, p. 110.

CHAPTER 11. Correlation.

The theory of correlation was first developed on definite assumptions as to the form of the distribution of frequency, the so-called "normal distribution" (Chap. 12) being assumed. Sir Francis Galton, in (242)–(244), developed the practical method, determining his coefficient (Galton's function, as it was termed at first) graphically. Edgeworth developed the theoretical side further in (240), and Pearson introduced the product-sum formula in (246)—both memoirs being written on the assumption of a "normal" distribution of frequency (cf. Chap. 12). The method used in this chapter is based on (247) and (248).

(236) BATEN, W. D., "Correction for the Moments of a Frequency-distribution in Two Variables," *Ann. Math. Stats.*, vol. 2, 1931, p. 309.

(237) BRAVAIS, A., "Analyse mathématique sur les probabilités des erreurs de situation d'un point," *Acad. des Sciences : Mémoires présentés par divers savants*, IIe série, t. 9, 1846, p. 255.

(238) DARBISHIRE, A. D., "Some Tables for illustrating Statistical Correlation," *Mem. and Proc. of the Manchester Lit. and Phil. Soc.*, vol. 51, 1907. (Tables and diagrams illustrating the meaning of values of the correlation coefficient from 0 to 1 by steps of a twelfth.)

(239) EDGEWORTH, F. Y., "On a New Method of reducing Observations relating to Several Quantities," *Phil. Mag.*, 5th Series, vol. 24, 1887, p. 222, and vol. 25, 1888, p. 184. (A method of treating correlated variables differing entirely from that described in this chapter, and based on the use of the median: the method involves the use of trial and error to some extent. For some illustrations see F. Y. Edgeworth and A. L. Bowley, *Jour. Roy. Stat. Soc.*, vol. 65, 1902, p. 341 *et seq.*)

(240) EDGEWORTH, F. Y., "On Correlated Averages," *Phil. Mag.*, 5th Series, vol. 34, 1892, p. 190.

(241) FRISCH, RAGNAR, "Correlation and Scatter in Statistical Variables," *Nordic Statistical Journal*, vol. 1, 1929, p. 36.

(242) GALTON, FRANCIS, "Regression towards Mediocrity in Hereditary Stature," *Jour. Anthrop. Inst.*, vol. 15, 1886, p. 246.

(243) GALTON, FRANCIS, "Family Likeness in Stature," *Proc. Roy. Soc.*, vol. 40, 1886, p. 42.

(244) GALTON, FRANCIS, "Correlations and their Measurement," *Proc. Roy. Soc.*, vol. 45, 1888, p. 135.

(245) PEARSON, KARL, "Notes on the History of Correlation," *Biometrika*, vol. 13, 1920, p. 25.

(246) PEARSON, KARL, "Regression, Heredity and Panmixia," *Phil. Trans. Roy. Soc.*, Series A, vol. 187, 1896, p. 253.

(247) YULE, G. U., "On the Significance of Bravais' Formulæ for Regression, etc., in the case of Skew Correlation," *Proc. Roy. Soc.*, vol. 60, 1897, p. 477.

(248) YULE, G. U., "On the Theory of Correlation," *Jour. Roy. Stat. Soc.*, vol. 60, 1897, p. 812.

CHAPTERS 12 AND 13. Normal Correlation and Further Theory of Correlation.

General.

(249) BRAVAIS, A., "Analyse mathématique sur les probabilités des erreurs de situation d'un point," *Acad. des Sciences : Mémoires présentés par divers savants*, IIe série, t. 9, 1846, p. 255.

(250) GALTON, FRANCIS, "Family Likeness in Stature," *Proc. Roy. Soc.*, vol. 40, 1886, p. 42.

(251) GALTON, FRANCIS, *Natural Inheritance*; Macmillan & Co., 1889.

(252) DICKSON, J. D. HAMILTON, Appendix to (250), *Proc. Roy. Soc.*, vol. 40, 1886, p. 63.

(253) EDGEWORTH, F. Y., "On Correlated Averages," *Phil. Mag.*, 5th Series, vol. 34, 1892, p. 190.

(254) PEARSON, KARL, "Regression, Heredity and Panmixia," *Phil. Trans. Roy. Soc.*, Series A, vol. 187, 1896, p. 253.

(255) PEARSON, KARL, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Phil. Mag.*, 6th Series, vol. 2, 1901, p. 559. (On the fitting of "principal axes" and the corresponding planes in the case of more than two variables.)

(256) PEARSON, KARL, "On the Influence of Natural Selection on the Variability and Correlation of Organs," *Phil. Trans. Roy. Soc.*, Series A, vol. 200, 1902, p. 1. (Based on the assumption of normal correlation.)

(257) PEARSON, KARL, and ALICE LEE, "On the Generalised Probable Error in Multiple Normal Correlation," *Biometrika*, vol. 6, 1908, p. 59.

(258) SHEPPARD, W. F., "On the Application of the Theory of Error to Cases of Normal Distribution and Normal Correlation," *Phil. Trans. Roy. Soc.*, Series A, vol. 192, 1898, p. 101.

(259) SHEPPARD, W. F., "On the Calculation of the Double-integral expressing Normal Correlation," *Cambridge Phil. Trans.*, vol. 19, 1900, p. 23.

(260) YULE, G. U., "On the Theory of Correlation," *Jour. Roy. Stat. Soc.*, vol. 60, 1897, p. 812.

(261) YULE, G. U., "On the Theory of Correlation for Any Number of Variables treated by a New System of Notation," *Proc. Roy. Soc.*, Series A, vol. 79, 1907, p. 182.

Applications to the Theory of Attributes, etc.

(262) PEARSON, KARL, "On the Correlation of Characters not Quantitatively Measurable," *Phil. Trans. Roy. Soc.*, Series A, vol. 195, 1900, p. 1. (*Cf.* criticism in ref. (80).)

(263) PEARSON, KARL, "On a New Method of Determining Correlation between a Measured Character A and a Character B of which only the Percentage of Cases wherein B exceeds (or falls short of) a Given Intensity is recorded for each grade of A," *Biometrika*, vol. 7, 1909, p. 96.

(264) PEARSON, KARL, "On a New Method of Determining Correlation, when one Variable is given by Alternative and the other by Multiple Categories," *Biometrika*, vol. 7, 1910, p. 248.

     See also the memoir (258) by Sheppard.

Various Methods and their Relation to Normal Correlation.

(265) PEARSON, KARL, "On the Theory of Contingency and its Relation to Association and Normal Correlation," *Drapers' Company Research Memoirs, Biometric Series I*; Dulau & Co., London, 1904.

(266) PEARSON, KARL, "On Further Methods of Determining Correlation," *Drapers' Company Research Memoirs, Biometric Series IV*. (Methods based on correlation of ranks: difference methods). Dulau & Co., London, 1907.

(267) PEARSON, KARL, and Others (editorial), "Tables for Determining the Volumes of a Bivariate Normal Surface," *Biometrika*, vol. 22, 1930, p. 1.

(268) SPEARMAN, C., "A Footrule for Measuring Correlation," *Brit. Jour. of Psychology*, vol. 2, 1906, p. 89. (The suggestion of a "rank" method: see Pearson's criticism and improved formula in (266), and Spearman's reply on some points in (269).)

(269) SPEARMAN, C., "Correlation Calculated from Faulty Data," *Brit. Jour. of Psychology*, vol. 3, 1910, p. 271.

(270) THORNDIKE, E. L., "Empirical Studies in the Theory of Measurement," *Archives of Psychology* (New York), 1907.

Fit of Regression Lines.

(271) FISHER, R. A., "The Goodness of Fit of Regression Formulæ, and the Distribution of Regression Coefficients," *Jour. Roy. Stat. Soc.*, vol. 85, 1922, p. 597.

(272) PEARSON, KARL, "On the Application of Goodness of Fit Tables to test Regression Curves and Theoretical Curves used to describe Observational or Experimental Data," *Biometrika*, vol. 11, 1916-17, p. 237.

Correlation in Case of Non-linear Regression.

(273) PEARSON, KARL, "On a General Method of Determining the Successive Terms in a Skew Regression Line," *Biometrika*, vol. 13, 1921, p. 296.
(274) PEARSON, KARL, "On the Correction necessary for the Correlation Ratio η,"
 • *Biometrika*, vol. 8, 1911, p. 254, and vol. 14, 1923, p. 412.
(275) PRETORIUS, S. J., "Skew Bivariate Frequency Surfaces, examined in the Light of Numerical Illustrations," *Biometrika*, vol. 22, 1930, p. 109.
(276) WICKSELL, S. D., "On Logarithmic Correlation, with an Application to the Distribution of Ages at First Marriage," *Meddelande fran Lunds Astronomiska Observatorium*, No. 84, 1917; Svenska Aktuarieforenings Tidskrift.
(277) WICKSELL, S. D., "The Correlation Function of Type A," *Kungl. Svenska Vetenskapsakademiens Handl.*, Bd. 58, 1917.

See also refs. (353)–(355), (377), (378) and (379).

CHAPTER 14. Partial Correlation.

(278) BROWN, J. W., M. GREENWOOD, and FRANCES WOOD, "A Study of Index-correlations," *Jour. Roy. Stat. Soc.*, vol. 77, 1914, pp. 317–346. (The partial or "solid" correlation ratio is used.)
(279) CAMP, BURTON H., "Mutually Consistent Multiple Regression Surfaces," *Biometrika*, vol. 17, 1925, p. 443.
(280) EDGEWORTH, F. Y., "On Correlated Averages," *Phil. Mag.*, 5th Series, vol. 34, 1892, p. 194.
(281) EZEKIEL, MORDECAI, "The Determination of Curvilinear Regression Surfaces in the Presence of Other Variables," *Jour. Amer. Stat. Assoc.*, vol. 21, 1926, p. 310.
(282) EZEKIEL, M., "The Application of the Theory of Error to Multiple and Curvilinear Correlation," *Jour. Amer. Stat. Assoc.*, vol. 24, 1929, Supplement, p. 99.
(283) HALL, PHILIP, "Multiple and Partial Correlation Coefficients in the case of an n-Fold Variate System," *Biometrika*, vol. 19, 1927, p. 100.
(284) HOOKER, R. H., and G. U. YULE, "Note on Estimating the Relative Influence of Two Variables upon a Third," *Jour. Roy. Stat. Soc.*, vol. 69, 1906, p. 197.
(285) HORST, P., "A General Method of Evaluating Multiple Regression Constants," *Jour. Amer. Stat. Assoc.*, vol. 27, 1932, p. 270.
(286) ISSERLIS, L., "On the Partial Correlation Ratio. Pt. I. Theoretical," *Biometrika*, vol. 10, 1914, pp. 391–411.
(287) ISSERLIS, L., "On the Partial Correlation Ratio. Pt. II. Numerical," *Biometrika*, vol. 11, 1916-17, p. 50.
(288) KELLEY, T. L., and F. S. SALISBURY, "An Iteration Method for determining Multiple Correlation Constants," *Jour. Amer. Stat. Assoc.*, vol. 21, 1926, p. 282.
(289) KELLEY, T. L., and Q. McNEMAR, "Doolittle *versus* the Kelley-Salisbury Iteration Method for Computing Multiple Regression Coefficients," *Jour. Amer. Stat. Assoc.*, vol. 24, 1929, p. 164.
(290) PEARSON, KARL, "Regression, Heredity and Panmixia," *Phil. Trans. Roy. Soc.*, Series A, vol. 187, 1896, p. 253.
(291) PEARSON, KARL, "On the Partial Correlation Ratio," *Proc. Roy. Soc.*, Series A, vol. 91, 1915, p. 492.
(292) ROMANOVSKY, V., "Sulle Regressione Multiple," *Giorn. dell' Ist. Ital. degli Attuari*, anno 2, 1931.
(293) TAPPAN, M., "On Partial Multiple Correlation Coefficients in a Universe of Manifold Characteristics," *Biometrika*, vol. 19, 1927, p. 39.
(294) THOMSON, G. H., "On the Computation of Regression Equations, Partial Correlations, etc.," *Brit. Jour. Psych.*, vol. 23, 1932, p. 64.
(295) TSCHUPROW, A. A., transl. by L. ISSERLIS, "The Mathematical Theory of the Statistical Methods employed in the Study of Correlation in the case of Three Variables," *Trans. Camb. Phil. Soc.*, vol. 23, 1928, p. 337.
(296) YULE, G. U., "On the Significance of Bravais' Formulæ for Regression, etc., in the case of Skew Correlation," *Proc. Roy. Soc.*, vol. 60, 1897, p. 477.
(297) YULE, G. U., "On the Theory of Correlation," *Jour. Roy. Stat. Soc.*, vol. 60, 1897, p. 812.
(298) YULE, G. U., "On the Theory of Correlation for Any Number of Variables treated by a New System of Notation," *Proc. Roy. Soc.*, Series A, vol. 79, 1907, p. 182.

Illustrative Applications of Economic Interest.

(299) HOOKER, R. H., "The Correlation of the Weather and the Crops," *Jour. Roy. Stat. Soc.*, vol. 65, 1907, p. 1.

(300) SNOW, E. C., "The Application of the Method of Multiple Correlation to the Estimation of Post-censal Populations," *Jour. Roy. Stat. Soc.*, vol. 74, 1911, p. 575.

(301) YULE, G. U., "An Investigation into the Causes of Changes in Pauperism in England, etc.," *Jour. Roy. Stat. Soc.*, vol. 62, 1899, p. 249.

CHAPTER 15. Correlation: Illustrations and Practical Methods.

(302) ANDERSON, OSKAR, *Die Korrelationsrechnung in der Konjunkturforschung* (Frankfurter Gesellschaft für Konjunkturforschung); Kurt Schroeder, Bonn, 1929.

(303) ANDERSON, O., "Nochmals über 'The Elimination of Spurious Correlation due to Position in Time or Space,'" *Biometrika*, vol. 10, 1914, pp. 269–279. (Detailed theory of the method discussed by "Student" in (327).)

(304) ANDERSON, O., "Ueber ein neues Verfahren bei Anwendung der 'Variate-difference' Methode," *Biometrika*, vol. 15, 1923, p. 134.

(305) ANDERSON, O., "Ueber die Anwendung der Differenzeumethode (Variate difference Method) bei Reihenausgleichungen, Stabilitätsuntersuchungen, und Korrelationsmessungen," *Biometrika*, vol. 18, 1926, p. 293.

(306) ANDERSON, O., "On the Logic of the Decomposition of Statistical Series into Separate Components," *Jour. Roy. Stat. Soc.*, vol. 90, 1927, p. 548.

(307) CAVE-BROWNE-CAVE, F. E., "On the Influence of the Time Factor on the Correlation between the Barometric Heights at Stations more than 1000 miles apart," *Proc. Roy. Soc.*, vol. 74, 1904, pp. 403–413.

(308) CAVE, BEATRICE M., and KARL PEARSON, "Numerical Illustrations of the Variate-difference Correlation Method," *Biometrika*, vol. 10, 1914, pp. 340–355.

(309) DARMOIS, G., "Analyse et comparaison des séries statistiques qui se développment dans le temps," *Metron*, vol. 8, Nos. 1–2, 1929, p. 211.

(310) FRISCH, RAGNAR, "A Method of Decomposing an Empirical Series into its Cyclical and Progressive Components," *Jour. Amer. Stat. Assoc.*, vol. 26, 1931, Supplement, p. 73.

(311) GUMBEL, E. J., "Spurious Correlation and its Significance in Physiology," *Jour. Amer. Stat. Assoc.*, vol. 21, 1926, p. 179.

(312) HARRIS, J. ARTHUR, "The Correlation between a Component, and between the Sum of Two or More Components, and the Sum of the Remaining Components of a Variable," *Quart. Pub. Amer. Stat. Assoc.*, vol. 15, 1917, p. 854.

(313) HERON, D., *On the Relation of Fertility in Man to Social Status*, "Drapers' Co. Research Memoirs: Studies in National Deterioration," I; Dulau & Co., London, 1906.

(314) HOOKER, R. H., "On the Correlation of the Marriage-rate with Trade," *Jour. Roy. Stat. Soc.*, vol. 64, 1901, p. 485.

(315) HOOKER, R. H., "On the Correlation of Successive Observations: illustrated by Corn Prices," *ibid.*, vol. 68, 1905, p. 696.

(316) HOOKER, R. H., "The Correlation of the Weather and the Crops," *ibid.*, vol. 70, 1907, p. 1.

(317) HOTELLING, H., "An Application of Analysis Situs to Statistics," *Bull. Amer. Math. Soc.*, July–August 1927, p. 467.

(318) JACOB, S. M., "On the Correlations of Areas of Matured Crops and the Rainfall," *Mem. Asiatic Soc. Bengal*, vol. 2, 1910, p. 847.

(319) JORDAN, CHARLES, "Sur la détermination de la tendance séculaire des grandeurs statistiques par la méthode des moindres carrés," *Jour. de la Société Hongroise de Statistique*, vol. 7, 1929, p. 567.

(320) MACAULAY, F. G., "Smoothing of Time Series," New York, National Bureau of Economic Research, 1931.

(321) MARCH, L., "Comparaison numérique de courbes statistiques," *Jour. de la Société de Statistique de Paris*, 1905, pp. 255 and 306.

(322) NORTON, J. P., *Statistical Studies in the New York Money Market*; Macmillan Co., New York, 1902. (Applications to financial statistics: an instantaneous average method, analogous to that of Example 15.5, is employed, but the instantaneous average is obtained by an interpolated logarithmic curve.)

(323) PEARSON, KARL, ALICE LEE, and L. BRAMLEY MOORE, "Genetic (reproductive) Selection: Inheritance of Fertility in Man and of Fecundity in Thoroughbred Racehorses," *Phil. Trans. Roy. Soc.*, Series A, vol. 192, 1899, p. 257.

(324) PEARSON, KARL, and E. M. ELDERTON, "On the Variate-difference Method," *Biometrika*, vol. 14, 1923, p. 281.

(325) SIPOS, ALEXANDER, "Practical Application of Jordan's Method for Trend Measurement;" Victor Hornyanszky Co., Ltd., Budapest, 1930.

(326) SMITH, B. B., "Combining the Advantages of First-difference and Deviation-from-Trend Methods of Correlating Time Series," *Jour. Amer. Stat. Assoc.*, vol. 21, 1926, p. 55.

(327) "STUDENT," "The Elimination of Spurious Correlation due to Position in Time or Space," *Biometrika*, vol. 10, 1914, pp. 179–180. (The extension of the difference method by the use of successive differences.)

(328) WICKSELL, S. D., "An Exact Formula for Spurious Correlation," *Metron*, vol. 1, No. 4, 1921, p. 33.

(329) WILL, HARRY S., "On Fitting Curves to Observational Series by the Method of Differences," *Ann. Math. Stats.*, vol. 1, 1930, p. 159.

(330) WORKING, H., and H. HOTELLING, "Applications of the Theory of Error to the Interpretation of Trends," *Jour. Amer. Stat. Assoc.*, vol. 24, 1929, Supplement, p. 73.

(331) YULE, G. U., "On the Time-correlation Problem," *Jour. Roy. Stat. Soc.*, vol. 84, 1921, p. 497.

(332) YULE, G. U., "Why do we sometimes get Nonsense Correlations between Time Series? A Study in Sampling and the Nature of Time Series," *Jour. Roy. Stat. Soc.*, vol. 89, 1926, p. 1.

(333) YULE, G. U., "On the Correlation of Total Pauperism with Proportion of Out-relief," *Economic Jour.*, vol. 5, 1895, p. 603, and vol. 6, 1896, p. 613.

(334) YULE, G. U., "An Investigation into the Causes of Changes in Pauperism in England chiefly during the last two Intercensal Decades," *Jour. Roy. Stat. Soc.*, vol. 62, 1899, p. 249.

(335) YULE, G. U., "On the Changes in the Marriage- and Birth-rates in England and Wales during the past Half-century, with an Inquiry as to their probable Causes," *Jour. Roy. Stat. Soc.*, vol. 69, 1906, p. 88.

CHAPTER 16. Miscellaneous Theorems Involving the Use of the Correlation Coefficient.

Effect of Errors of Observation on the Correlation Coefficient.

(336) BROWN, W., "Some Experimental Results in Correlation," *Proceedings of the Sixth International Congress of Psychology, Geneva*, August 1909.

(337) HART, BERNARD, and C. SPEARMAN, "General Ability, its Existence and Nature," *Brit. Jour. Psych.*, vol. 5, 1912, p. 51. (For controversy about these formulæ, cf. ref. (14), Brown and Thomson, and references there given, critical notice in *Brit. Jour. Psych.*, vol. 12, 1921, p. 100, and also (342) below.)

(338) JACOB, S. M., "On the Correlations of Areas of Matured Crops and the Rainfall," *Mem. Asiatic Soc. Bengal*, vol. 2, 1910, p. 847. (§ 7 contains remarks on the effects of errors on the correlations and regressions, with especial reference to this problem.)

(339) SPEARMAN, C., "The Proof and Measurement of Association between Two Things," *Amer. Jour. Psych.*, vol. 15, 1904, p. 88.

(340) SPEARMAN, C., "Demonstration of Formulæ for True Measurement of Correlation," *Amer. Jour. Psych.*, vol. 18, 1907, p. 161.

(341) SPEARMAN, C., "Correlation Calculated from Faulty Data," *Brit. Jour. Psych.*, vol. 3, 1910, p. 271.

(342) STEAD, H. G., "The Correction of Correlation Coefficients," *Jour. Roy. Stat. Soc.*, vol. 86, 1923, p. 412.

Correlations between Indices, etc.

(343) BROWN, J. W., M. GREENWOOD, and FRANCES WOOD, "A Study of Index-correlations," *Jour. Roy. Stat. Soc.*, vol. 77, 1914, pp. 317–46.

(344) GALTON, FRANCIS, "Note to the Memoir by Prof. Karl Pearson on Spurious Correlation," *Proc. Roy. Soc.*, vol. 60, 1897, p. 498. (See (345) overleaf.)

(345) PEARSON, KARL, "On a Form of Spurious Correlation which may arise when Indices are used in the Measurement of Organs," *Proc. Roy. Soc.*, vol. 60, 1897, p. 489. (§§ 8, 9.)

(346) YULE, G. U., "On the Interpretation of Correlations between Indices or Ratios," *Jour. Roy. Stat. Soc.*, vol. 73, 1910, p. 644.

The Weighted Mean.

(347) PEARSON, KARL, "Note on Reproductive Selection," *Proc. Roy. Soc.*, vol. 59, 1896, p. 301.

Standardisation or Correction of Death-rates, etc.

For the methods of standardisation in present use in England and Wales, see *Seventy-fourth Annual Report of the Registrar-General of England and Wales, 1911*, Cd. 6578, 1913.

Papers (349) and (351) suggested methods of standardising the birth-rate.

(348) HERON, DAVID, "The Influence of Defective Physique and Unfavourable Home Environment on the Intelligence of School-children," *Eugenics Laboratory Memoirs*, 8; Dulau & Co., London, 1910.

(349) NEWSHOLME, A., and T. H. C. STEVENSON, "The Decline of Human Fertility in the United Kingdom and other Countries, as shown by Corrected Birth-rates," *Jour. Roy. Stat. Soc.*, vol. 69, 1906, p. 34.

(350) WOLFENDEN, H. H., "On the Methods of Comparing the Mortalities of Two or More Communities, and the Standardisation of Death-rates," *Jour. Roy. Stat. Soc.*, vol. 88, 1923, p. 399.

(351) YULE, G. U., "On the Changes in the Marriage- and Birth-rates in England and Wales during the past Half-century, etc.," *Jour. Roy. Stat. Soc.*, vol. 69, 1906, p. 88.

(352) YULE, G. U., "On Some Points Relating to Vital Statistics, more especially Statistics of Occupational Mortality," *Jour. Roy. Stat. Soc.*, vol. 97, 1934, p. 1. (Contains a full discussion of methods of standardisation.)

Theory of Correlation in the case of Non-linear Regression.

See refs. (273)–(277) and the following:—

(353) BLAKEMAN, J., "On Tests for Linearity of Regression in Frequency-distributions," *Biometrika*, vol. 4, 1905, p. 332.

(354) PEARSON, KARL, *On the General Theory of Skew Correlation and Non-linear Regression*, "Drapers' Co. Research Memoirs: Biometric Series," II; Dulau & Co., London, 1905. (The "correlation ratio.")

(355) PEARSON, KARL, "On a Correction to be made to the Correlation Ratio," *Biometrika*, vol. 8, 1911, p. 254, and vol. 14, 1923, p. 412.

Abbreviated Methods of Calculation.

(356) HARRIS, J. ARTHUR, "A Short Method of Calculating the Coefficient of Correlation in the case of Integral Variates," *Biometrika*, vol. 7, 1909, p. 214. (Not an approximation, but a true short method.)

(357) HARRIS, J. ARTHUR, "On the Calculation of Intra-class and Inter-class Coefficients of Correlation from Class-moments when the Number of possible Combinations is large," *Biometrika*, vol. 9, 1914, pp. 446–472.

CHAPTER 17.   Simple Curve Fitting.

See refs. (319), (329) of Chapter 15, and the following:—

(358) AITKEN, A. C., "On the Graduation of Data by the Orthogonal Polynomials of Least Squares," *Proc. Roy. Soc. Edin.*, vol. 53, 1933, p. 54.

(359) AITKEN, A. C., "On Fitting Polynomials to Weighted Data by Least Squares," *Proc. Roy. Soc. Edin.*, vol. 54, 1933, p. 1; and "On Fitting Polynomials to Data with Weighted and Correlated Errors," *Proc. Roy. Soc. Edin.*, vol. 54, 1933, p. 12.

(360) AITKEN, A. C., "On the Orthogonal Polynomials in Frequencies of Type B," *Proc. Roy. Soc. Edin.*, vol. 52, 1932, p. 174.

(361) AITKEN, A. C., and A. OPPENHEIM, "On Charlier's New Form of the Frequency Function," *Proc. Roy. Soc. Edin.*, vol. 51, 1931, p. 35.

(362) ALLAN, F. E., "The General Form of the Orthogonal Polynomials for Simple Series, with Proofs of their Simple Properties," *Proc. Roy. Soc. Edin.*, vol. 50, 1930, p. 310.

(363) BIRGE, R. T., and J. D. SHEA, "A Rapid Method of Calculating the Least Squares Solution of a Polynomial of Any Degree," *Univ. of California Pub. in Maths.*, vol. 2, 1927, p. 67.

(364) CHOTIMSKY, V., *The Smoothing of Statistical Series by Least Squares (Tschebycheff's Method)*. (In Russian.) Soviet Press, Moscow and Leningrad, 1925.

(365) CONDON, E., "The Rapid Fitting of a Certain Class of Empirical Formulæ by the Method of Least Squares," *Univ. of California Pub. in Maths.*, vol. 2, 1927, p. 55.

(366) DAVIS, H. T., "Polynomial Approximation by the Method of Least Squares," *Ann. Math. Stats.*, vol. 4, 1933, p. 155.

(367) DAVIS, H. T., and V. V. LATSHAW, "Formulæ for the Fitting of Polynomials to Data by the Method of Least Squares," *Ann. Math.* (2nd Series), vol. 31, 1930, No. 1, p. 52.

(368) FISHER, R. A., "Studies in Crop Variation: I," *Jour. Agricultural Science*, vol. 11, 1921, p. 107.

(369) GINI, C., "Sull' interpolazione di una retta quando i valori della variabile indipendente sono affetti dà errori accidentali," *Metron*, vol. 1, 1922, part 3, p. 53.

(370) GINI, C., "Considerazioni sull' interpolazione e la perequazione delle serie statistiche," *Metron*, vol. 1, 1922, part 3, p. 3.

(371) GRAM, J. P., "Om rackkendviklinger bestemte ved Hjaelp af de mindste Kvadraters Methode," 1879, Copenhagen. Reprinted as "Über die Entwicklung realer Functionen in Reihen mittelst der Methode der Kleinsten Quadraten," *Jour. für Math.*, vol. 94, 1894, p. 41.

(372) GREENLEAF, H. E. H., "Curve Approximation by Means of Functions Analogous to the Hermite Polynomials," *Ann. Math. Stats.*, vol. 3, 1932, p. 204. (Contains references.)

(373) HENDRICKS, W. A., "The Use of the Relative Residual in the Application of the Method of Least Squares," *Ann. Math. Stats.*, vol. 2, 1931, p. 458.

(374) ISSERLIS, L., "Note on Tchebycheff's Interpolation Formula," *Biometrika*, vol. 19, 1927, p. 87.

(375) JORDAN, CH., *Statistique mathématique*; Gauthier-Villars, Paris, 1927.

(376) JORDAN, CH., "Approximation and Graduation according to the Principle of Least Squares by Orthogonal Polynomials," *Ann. Math. Stats.*, vol. 3, 1932, p. 257.

(377) PEARSON, KARL, "On the Systematic Fitting of Curves to Observations and Measurements," *Biometrika*, vol. 1, 1901, p. 265, and vol. 2, 1902, p. 1.

(378) PEARSON, KARL, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Phil. Mag.*, 6th Series, vol. 2, 1901, p. 559.

(379) PEARSON, KARL, "On a General Theory of the Method of False Position," *Phil. Mag.*, 6th Series, vol. 4, 1903.

(380) PEARSON, KARL, "On a General Method of Determining the Successive Terms in a Skew Regression Line," *Biometrika*, vol. 13, 1921, p. 296.

(381) PIETRA, G., "Interpolating Plane Curves," *Metron*, vol. 3, 1924, p. 311.

(382) REED, L. F., "Fitting Straight Lines," *Metron*, vol. 1, 1922, part 3, p. 54.

(383) RHODES, E. C., "On the Fitting of Parabolic Curves to Statistical Data," *Jour. Roy. Stat. Soc.*, vol. 93, 1930, p. 569.

(384) ROMANOVSKY, V., "Note on Orthogonalising Series of Functions and Interpolation," *Biometrika*, vol. 19, 1927, p. 93.

(385) SNOW, E. C., "On Restricted Lines and Planes of Closest Fit to Systems of Points in Any Number of Dimensions," *Phil. Mag.*, 6th Series, vol. 21, 1911, p. 367.

(386) TSCHEBYCHEFF, P. L. See numerous papers in his collected works, *Œuvres.*

(387) WHITTAKER and ROBINSON, *Calculus of Observations*; Blackie & Son, London, 2nd Ed., 1932.

CHAPTER 18. Preliminary Notions on Sampling.

Theory of Probability and its Applications to Statistics.

(388) KEYNES, J. M., *A Treatise on Probability*; Macmillan, London, 1921.
(389) POINCARÉ, H., *Calcul des Probabilités*; Gauthier-Villars, Paris, 1896.
(390) VENN, J. A., *The Logic of Chance*; Macmillan, London, 3rd Ed., 1888.

      (388) and (390) treat of probability from the point of view of its logical and philosophical foundations, and give a useful general introduction to the subject. See also refs. (7) and (9).

Bias in Sampling.

(391) KISER, C. V., "Pitfalls in Sampling for Population Study," *Jour. Amer. Stat. Assoc.*, vol. 29, 1934, pp. 250-256.
(392) YATES, F., "Some Examples of Biased Sampling," *Annals of Eugenics*, vol. 6, 1935, pp. 202-213.

Various Sampling Methods.

(393) BOWLEY, A. L., "Working-class Households in Reading," *Jour. Roy. Stat. Soc.*, vol. 76, 1913, p. 672.
(394) BOWLEY, A. L., "Measurement of the Precision attained in Sampling," *Bull. Int. Stat. Inst.*, vol. 22, 1er livre.
(394a) HILTON, JOHN, "Enquiry by Sample; an Experiment and its Results," *Jour. Roy. Stat. Soc.*, vol. 87, 1924, p. 544.
(395) JENSEN, A., "Report on the Representative Method in Statistics," *Bull. Int. Stat. Inst.*, vol. 22, 1er livre.
(396) JENSEN, A., "Purposive Selection," *Jour. Roy. Stat. Soc.*, vol. 91, 1928, pp. 541-547.
(397) NEYMAN, J., "On Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection," *Jour. Roy. Stat. Soc.*, vol. 97, 1934, pp. 558-625.

CHAPTER 19. Sampling of Attributes—Large Samples.

(Including references to experimental results of dice-throwing, etc.)

(398) DARBISHIRE, A. D., "Some Tables for Illustrating Statistical Correlation," *Mem. and Proc. of the Manchester Lit. and Phil. Soc.*, vol. 51, 1907.
(399) DETLEFSEN, J. A., "Fluctuations of Sampling in a Mendelian Population," *Genetics*, vol. 3, 1918, p. 599.
(400) EDGEWORTH, F. Y., "Miscellaneous Applications of the Calculus of Probabilities," *Jour. Roy. Stat. Soc.*, vols. 60, 61, 1897-98 (especially part 2, vol. 61, p. 119).
(401) EDGEWORTH, F. Y., Article on the "Law of Error" in the Tenth Edition of the *Encyclopædia Britannica*, vol. 28, 1902, p. 280; or on "Probability," Eleventh Edition, vol. 22 (especially Part 2, pp. 390 *et seq.*).
(402) EDGEWORTH, F. Y., "Methods of Statistics," *Jour. Roy. Stat. Soc.*, jubilee volume, 1885, p. 181.
(403) GREENWOOD, M., "On Errors of Random Sampling in certain Cases not suitable for the Application of a 'Normal Curve of Frequency,'" *Biometrika*, vol. 9, 1913, pp. 69-90. (If an event has succeeded p times in n trials, what are the chances of 0, 1, . . . m successes in m subsequent trials? Tables for small samples.)
(404) LEXIS, W., *Zur Theorie der Massenerscheinungen in der menschlichen Gesellschaft*; Freiburg, 1877.
(405) LEXIS, W., *Abhandlungen zur Theorie der Bevölkerungs und Moralstatistik*; Fischer, Jena, 1903. (Contains, with new matter, reprints of some of Professor Lexis' earlier papers in a form convenient for reference.)
(405a) PARKES, A. S., "Studies on the Sex Ratio and Related Phenomena," *Biometrika*, vol. 15, 1923, p. 373.
(406) PEARSON, KARL, "Skew Variation in Homogeneous Material," *Phil. Trans. Roy. Soc.*, Series A, vol. 186, 1895, p. 343. (Sections 2 to 6 on the binomial distribution.)

(407) PEARSON, KARL, "On certain Properties of the Hypergeometrical Series, and on the fitting of such Series to Observation Polygons in the Theory of Chance," *Phil. Mag.*, 5th Series, vol. 47, 1899, p. 236. (An expansion of one section of ref. (406), dealing with the problem of drawing samples from a bag containing a limited number of white and black balls, from the standpoint of the frequency-distribution of the number of white or black balls in the samples.)

(408) PEARSON, KARL, "On the Difference and the Doublet Tests for Ascertaining whether Two Samples have been drawn from the Same Population," *Biometrika*, vol. 16, 1924, p. 249.

(409) POISSON, S. D., "Sur la proportion des naissances des filles et des garçons," *Mémoires de l'Acad. des Sciences*, vol. 9, 1829, p. 239. (Principally theoretical: the statistical illustrations very slight.)

(410) RHODES, E. C., "On the Problem whether Two Given Samples can be supposed to have been drawn from the Same Population," *Biometrika*, vol. 16, 1924, p. 239, and *Metron*, vol. 5, 1925, p. 3.

(411) VENN, JOHN, *The Logic of Chance*, 3rd Ed.; Macmillan, London, 1888.

(412) VIGOR, H. D., and G. U. YULE, "On the Sex Ratios of Births in the Registration Districts of England and Wales, 1881–90," *Jour. Roy. Stat. Soc.*, vol. 69, 1906, p. 576.

(413) WESTERGAARD, H., *Die Grundzüge der Theorie der Statistik*; Fischer, Jena, 1890, and 2nd Ed., enlarged, with H. C. NYBØLLE, 1928.

(414) YULE, G. U., "Fluctuations of Sampling in Mendelian Ratios," *Proc. Camb. Phil. Soc.*, vol. 17, 1914, p. 425.

See also under Binomial, Normal Curve, Chapter 10, and the General References for Standard Errors below, Chapters 20–21.

CHAPTERS 20 AND 21. Sampling of Variables – Large Samples.

The probable errors of various special coefficients, etc., are generally dealt with in the memoirs concerning them, reference to which has been made in the lists of previous chapters : reference has also been made before to most of the memoirs concerning errors of sampling in proportions or percentages. The following is a classification of some of the memoirs in the list below :—

General: (415), (421), (422), (424), (425), (426), (429), (431), (437), (444), (447), (452), (453), (455), (459), (460), (468).

Averages and percentiles: (416), (427), (428), (430), (436), (442), (445), (446), (475), (482), (483).

Standard deviation: (423), (428), (432), (449), (454), (470), (475).

Coefficient of correlation (product-sum and partial correlations): (417), (428), (434), (435), (441), (457), (470), (478), (479), (490), (491).

Coefficient of correlation, other methods, etc.: (418), (443), (460), (465), (467), (481), (487).

Coefficients of association: (491).

Coefficient of contingency: (419), (448), (466), (489).

Moments: (437), (480), (484), (485), (486), (488).

Coefficient of variation: (451).

(415) BAKER, GEORGE A., "Random Samples from Non-homogeneous Populations," *Metron*, vol. 8, No. 3, 1930, p. 67.

(416) BAKER, GEORGE A., "Distribution of the Means of Samples of n drawn at random from a Population represented by a Gram-Charlier Series," *Ann. Math. Stats.*, vol. 1, 1930, p. 199, and note by C. C. CRAIG, *ibid.*, vol. 2, 1931, p. 99.

(417) BISPHAM, J. W., "An Experimental Determination of the Distribution of the Partial Correlation Coefficient in Samples of Thirty," *Proc. Roy. Soc.*, A, vol. 97, 1920, and *Metron*, vol. 2, 1923, p. 684.

(418) BLAKEMAN, J., "On Tests for Linearity of Regression in Frequency-distributions," *Biometrika*, vol. 4, 1905, p. 332.

(419) BLAKEMAN, J., and KARL PEARSON, "On the Probable Error of the Coefficient of Mean Square Contingency," *Biometrika*, vol. 5, 1906, p. 191.

(420) BORTKIEWICZ, L. VON, "The Relation between Stability and Homogeneity," *Ann. Math. Stats.*, vol. 2, 1931, p. 1.

(421) BOWLEY, A. L., *The Measurement of Groups and Series*; C. & E. Layton, London, 1903.

(422) CARVER, H. C., "Fundamentals of the Theory of Sampling," *Ann. Math. Stats.*, vol. 1, 1930, pp. 101 and 205.

(423) CARVER, H. C., "The Interdependence of Sampling and Frequency-distribution Theory," *Ann. Math. Stats.*, vol. 2, 1931, p. 82.

(424) CRAIG, C. C., "An Application of Thiele's Seminvariants to the Sampling Problem," *Metron*, vol. 7, 1928, p. 3.

(425) CRAIG, C. C., "Sampling in the case of Correlated Observations," *Ann. Math. Stats.*, vol. 2, 1931, p. 324.

(426) CRAIG, C. C., "Note on the Distribution of Samples of N drawn from a Type A Population," *Ann. Math. Stats.*, vol. 2, 1931, p. 99.

(427) DODD, E. L., "The Probability of the Arithmetic Mean compared with that of certain other Functions of the Measurements," *Ann. Maths.*, vol. 14, 1912-13.

(428) DUNLAP, H. F., "An Empirical Determination of the Distribution of Means, Standard Deviations and Correlation Coefficients drawn from Rectangular Populations," *Ann. Math. Stats.*, vol. 2, 1931, p. 66.

(428a) EDGEWORTH, F. Y., "Observations and Statistics: An Essay on the Theory of Errors of Observation and the First Principles of Statistics," *Cambridge Phil. Trans.*, vol. 14, 1885, p. 139.

(429) EDGEWORTH, F. Y., "Problems in Probabilities," *Phil. Mag.*, 5th Series, vol. 22, 1886, p. 371.

(430) EDGEWORTH, F. Y., "The Choice of Means," *Phil. Mag.*, 5th Series, vol. 24, 1887, p. 268.

(431) EDGEWORTH, F. Y., "On the Probable Errors of Frequency Constants," *Jour. Roy. Stat. Soc.*, vol. 71, 1908, pp. 381, 499, 651; and Addendum, vol. 72, 1909, p. 81.

(432) FELDMAN, H. M., "The Distribution of the Precision Constant and its Square in Samples from a Normal Population," *Ann. Math. Stats.*, vol. 3, 1932, p. 20.

(433) FIELLER, E. C., "The Distribution of the Index in a Normal Bivariate Population," *Biometrika*, vol. 24, 1932, p. 428.

(434) FISHER, R. A., "The Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population," *Biometrika*, vol. 10, 1915, p. 507.

(435) FISHER, R. A., "The Distribution of the Partial Correlation Coefficient," *Metron*, vol. 3, 1924, p. 329.

(436) FISHER, R. A., "A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error and the Mean Square Error," *Monthly Notices, Royal Astr. Soc.*, vol. 80, 1920, p. 75.

(437) FISHER, R. A., "Moments and Product-moments of Sampling Distributions," *Proc. Lond. Math. Soc.*, Series 2, vol. 30, 1928, p. 199.

(438) FISHER, R. A., "The Moments of the Distribution for Normal Samples of Measures of Departure from Normality," *Proc. Roy. Soc.*, A, vol. 130, 1930, p. 16.

(439) GIBSON, WINIFRED, "Tables for Facilitating the Computation of Probable Errors," *Biometrika*, vol. 4, 1906, p. 385.

(440) HERON, D., "An Abac to determine the Probable Errors of Correlation Coefficients," *Biometrika*, vol. 7, 1910, p. 411. (A diagram giving the probable error for any number of observations up to 1000.)

(441) HERON, D., "On the Probable Error of a Partial Correlation Coefficient," *Biometrika*, vol. 7, 1910, p. 411.

(442) HOJO, T., "Distribution of the Median, Quartiles and Interquartile Distance in Samples from a Normal Population," *Biometrika*, vol. 23, 1931, p. 315.

(443) HOLZINGER, K. S., and A. E. R. CHURCH, "On the Means of Samples from a U-shaped Population," *Biometrika*, vol. 20A, 1929, p. 361.

(444) HOTELLING, HAROLD, "The Distribution of Correlation Ratios Calculated from Random Data," *Proc. Nat. Acad. Sci.*, vol. 11, 1925, p. 657.

(445) HOTELLING, H., "The Consistency and Ultimate Distribution of Optimum Statistics," *Trans. Amer. Math. Soc.*, vol. 32, 1930, p. 847.

(445a) IRWIN, J. O., "On the Frequency-distribution of the Means of Samples from a Population having Any Law of Frequency with Finite Moments, etc.," *Biometrika*, vol. 19, 1927, p. 225, and vol. 22, 1929, p. 431.

(446) IRWIN, J. O., "On the Frequency-distribution of the Means of Samples from Populations of certain of Pearson's Types," *Metron*, vol. 7, No. 4, 1930, p. 51.

(447) ISSERLIS, L., "On the Conditions under which the 'Probable Errors' of Frequency-distributions have a real Significance," *Proc. Roy. Soc.*, Series A, vol. 92, 1915, p. 23.

(448) KONDO, T., "On the Standard Error of the Mean Square Contingency," *Biometrika*, vol. 21, 1929, p. 376.

(449) KONDO, T., "A Theory of the Sampling Distribution of Standard Deviations," *Biometrika*, vol. 22, 1930, p. 36.

(450) LAPLACE, PIERRE SIMON, Marquis de, *Théorie des probabilités*, 2e édn., 1814. (With four supplements.)

(451) McKAY, A. T., "The Distribution of the Estimated Coefficient of Variation," *Jour. Roy. Stat. Soc.*, vol. 94, 1931, p. 564.

(452) MEIDELL, H. BIRGER, "Sur la probabilité des erreurs," *Comptes rendus*, vol. 176, 1923, p. 280.

(453) PEARL, RAYMOND, "The Calculation of Probable Errors of Certain Constants of the Normal Curve," *Biometrika*, vol. 5, 1906, p. 190.

(454) PEARL, RAYMOND, "On certain Points concerning the Probable Error of the Standard Deviation," *Biometrika*, vol. 6, 1908, p. 112. (On the amount of divergence, in certain cases, from the standard error $\sigma/\sqrt{2n}$ in the case of a normal distribution.)

(455) PEARSON, EGON S., "A Further Development of Tests for Normality," *Biometrika*, vol. 22, 1930, p. 239.

(456) PEARSON, E. S., "The Probable Error of a Class-index Correlation," *Biometrika*, vol. 14, 1923, p. 261.

(457) PEARSON, E. S., "Note on the Approximations to the Probable Error of a Coefficient of Correlation," *Biometrika*, vol. 16, 1924, p. 196.

(458) PEARSON, E. S., "The Percentage Limits for the Distribution of Range in Samples from a Normal Population," *Biometrika*, vol. 24, 1932, p. 404.

(459) PEARSON, KARL, and L. N. G. FILON, "On the Probable Errors of Frequency Constants, and on the Influence of Random Selection on Variation and Correlation," *Phil. Trans. Roy. Soc.*, Series A, vol. 191, 1898, p. 229.

(460) PEARSON, KARL (editorial), "On the Probable Errors of Frequency Constants, Part 1," *Biometrika*, vol. 2, 1903, p. 273, "Part 2," *ibid.*, vol. 9, 1913, p. 1, and "Part 3," *ibid.*, vol. 13, 1920, p. 113. (Useful for the general formulæ given, based on the general case without respect to the form of the frequency-distribution.)

(461) PEARSON, KARL, "On the Criterion that a Given System of Deviations from the Probable in the case of a Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling," *Phil. Mag.*, vol. 50, Series 5, 1900, p. 157.

(462) PEARSON, KARL, "On the Curves which are most suitable for describing the Frequency of Random Samples of a Population," *Biometrika*, vol. 5, 1906, p. 172.

(463) PEARSON, KARL, "Note on the Significant or Non-significant Character of a Sub-sample drawn from a Sample," *Biometrika*, vol. 5, 1906, p. 181.

(464) PEARSON, KARL, "On the Probability that two Independent Distributions of Frequency are really Samples from the same Population," *Biometrika*, vol. 8, 1911, p. 250, and vol. 10, 1914, p. 85.

(465) PEARSON, KARL, "On the Probable Error of a Coefficient of Correlation as found from a Fourfold Table," *Biometrika*, vol. 9, 1913, p. 22.

(466) PEARSON, KARL, "On the Probable Error of a Coefficient of Mean Square Contingency," *Biometrika*, vol. 10, 1915, p. 590.

(467) PEARSON, KARL, "On the Probable Error of Biserial η," *Biometrika*, vol. 11, 1915–17, p. 292.

(468) PEARSON, KARL, and BRENDA STOESSIGER, "Tables of the Probability Integrals of Symmetrical Frequency-curves in the case of Low Powers, such as arise in the Theory of Small Samples," *Biometrika*, vol. 22, 1931, p. 253.

(469) PEARSON, KARL, "On the Nature of the Relationship between Two of 'Student's' Variates (z_1 and z_2) when Samples are taken from a Bivariate Normal Population," *Biometrika*, vol. 22, 1931, p. 405.

(470) PEARSON, KARL, "Historical Note on the Distribution of Standard Deviations of Samples of Any Size from an Indefinitely Large Normal Parent Population," *Biometrika*, vol. 23, 1931, p. 416.

(471) PEPPER, JOSEPH, "Studies in the Theory of Sampling," *Biometrika*, vol. 21, 1929, p. 231.

(472) PEPPER, JOSEPH, "The Sampling Distribution of the Third Moment Coefficient: An Experiment," *Biometrika*, vol. 24, 1932, p. 55.

(473) RHIND, A., "Tables for Facilitating the Computation of Probable Errors of the Chief Constants of Skew Frequency-distributions," *Biometrika*, vol. 7, 1909–10, pp. 127 and 386.

(474) RHODES, E. C., "The Comparison of Two Sets of Observations," *Jour. Roy. Stat. Soc.*, vol. 89, 1926, p. 544.

(475) RHODES, E. C., "The Precision of Means and Standard Deviations when the Individual Errors are Correlated," *Jour. Roy. Stat. Soc.*, vol. 90, 1927, p. 135.

(476) ST GEORGESCU, N., "Further Contributions to the Sampling Problem," *Biometrika*, vol. 24, 1932, p. 65.

(477) SHEPPARD, W. F., "On the Application of the Theory of Error to Cases of Normal Distribution and Normal Correlation," *Phil. Trans. Roy. Soc.*, Series A, vol. 192, 1898, p. 101.

(478) SOPER, H. E., "On the Probable Error of the Correlation Coefficient to a Second Approximation," *Biometrika*, vol. 9, 1913, p. 91.

(479) SOPER, H. E., "On the Probable Error of the Bi-serial Expression for the Correlation Coefficient," *Biometrika*, vol. 10, 1914, p. 384.

(480) SOPER, H. E., "Sampling Moments of Moments of Samples of n Units each drawn from an Unchanging Sampled Population, from the Point of View of Semi-invariants," *Jour. Roy. Stat. Soc.*, vol. 93, 1930, p. 104.

(481) "STUDENT," "An Experimental Determination of the Probable Error of Dr. Spearman's Correlation Coefficients," *Biometrika*, vol. 13, 1921, p. 263.

(482) "STUDENT," "On the Distribution of Means of Samples which are not drawn at Random," *Biometrika*, vol. 7, 1909, p. 210.

(483) TCHEBYCHEFF, P. L. DE, "Des valeurs moyennes," *Jour. de Maths.* (2), vol. 12, 1867, pp. 177–184.

(484) TSCHUPROW, A. A., "On the Mathematical Expectation of the Moments of Frequency-distributions," *Biometrika*, vol. 12, 1918–19, pp. 140 and 185, and vol. 13, 1921, p. 283; and *Metron*, vol. 2, 1923, pp. 461 and 646.

(485) WISHART, J., "The Derivation of certain High-order Sampling Product-moments from a Normal Population," *Biometrika*, vol. 22, 1930, p. 224.

(486) WISHART, J., "Notes on Frequency Constants," *Jour. Inst. of Actuaries*, vol. 62, 1931, p. 174.

(487) WISHART, J., "The Mean and Second-moment Coefficient of the Multiple Correlation Coefficient in Samples from a Normal Population," *Biometrika*, vol. 22, 1931, p. 353. (With an editorial appendix of tables of the mean value and squared standard deviation of a multiple correlation coefficient.)

(488) WISHART, J., and M. S. BARTLETT, "The Distribution of Second-order Moment Coefficients in Small Samples," *Proc. Camb. Phil. Soc.*, vol. 28, 1932, p. 455.

    On the problem of fluctuations of sampling in correlations between time-series, see also YULE (332).

(489) YOUNG, ANDREW, and KARL PEARSON, "On the Probable Error of a Coefficient of Contingency without Approximation," *Biometrika*, vol. 11, 1916–17, p. 215.

(490) YULE, G. U., "On the Theory of Correlation for Any Number of Variables treated by a New System of Notation," *Proc. Roy. Soc.*, Series A, vol. 79, 1907, p. 182. (See pp. 192–193 at end.)

(491) YULE, G. U., "On the Methods of Measuring Association between Two Attributes," *Jour. Roy. Stat. Soc.*, vol. 75, 1912. (Probable error of the correlation coefficient for a fourfold table, of association coefficients, etc.)

    Reference may also be made to the following, which deal for the most part with the effects of errors other than errors of sampling:—

(492) BOWLEY, A. L., "Relations between the Accuracy of an Average and that of its Constituent Parts," *Jour. Roy. Stat. Soc.*, vol. 60, 1897, p. 855.

(493) BOWLEY, A. L., "The Measurement of the Accuracy of an Average," *Jour. Roy. Stat. Soc.*, vol. 75, 1911, p. 77.

CHAPTER 22.   The χ^2 Distribution.

(494) BOWLEY, A. L., and R. L. CONNOR, "Tests of Correspondence between Statistical Grouping and Formulæ," *Economica*, 1923, p. 1.

(495) FISHER, R. A., "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P," *Jour. Roy. Stat. Soc.*, vol. 85, 1922, p. 87.

(496) FISHER, R. A., "On the Mathematical Foundations of Theoretical Statistics," *Phil. Trans.*, Series A, vol. 222, 1922, pp. 309-368.

(497) FISHER, R. A., "The Conditions under which χ^2 measures the Discrepancy between Observation and Hypothesis," *Jour. Roy. Stat. Soc.*, vol. 87, 1924, p. 442.

(498) FISHER, R. A., "Statistical Tests of Agreement between Observation and Hypothesis" (with a note in reply by A. L. Bowley), *Economica*, 1923, p. 139.

(499) IRWIN, J. O., "Note on the χ^2 Test for Goodness of Fit," *Jour. Roy. Stat. Soc.*, vol. 92, 1929, p. 264.

(500) NEYMAN, J., and E. S. PEARSON, "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," *Biometrika*, vol. 20A, 1928, pp. 175 and 263.

(501) NEYMAN, J., and E. S. PEARSON, "Further Notes on the χ^2 Distribution," *Biometrika*, vol. 22, 1931, pp. 298-305.

(502) PEARSON, KARL, "On the Criterion that a Given System of Deviations from the Probable in the case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling," *Phil. Mag.*, vol. 50, Series 5, 1900, pp. 157-175.

(503) PEARSON, KARL, "Multiple Cases of Disease in the Same House," *Biometrika*, vol. 9, 1913, p. 28. (A modification of the goodness of fit test to cover such statistics as those indicated by the title.)

(504) PEARSON, KARL, "On the Application of Goodness of Fit Tables to Test Regression Curves and Theoretical Curves to Describe Observational or Experimental Data," *Biometrika*, vol. 11, 1915, p. 239.

(505) PEARSON, KARL, "On a Brief Proof of the Fundamental Formula for Testing the Goodness of Fit of Frequency-distribution and on the Probable Error of P," *Phil. Mag.*, vol. 30D (6th Series), 1916, p. 369.

(506) PEARSON, KARL, "On the χ^2 Test of Goodness of Fit," *Biometrika*, vol. 14, 1922, p. 186; and "Further Note," *ibid.*, p. 418.

(507) PEARSON, KARL, "Note on the Relation of the (P, χ^2) Test to the Distribution of Standard Deviations in Samples from a Normal Population," *Biometrika*, vol. 19, 1927, p. 215.

(508) PEARSON, KARL, "Experimental Discussion of the Test for Goodness of Fit," *Biometrika*, vol. 24, 1932, pp. 351-381.

(509) ROBINSON, SELBY, "An Experiment regarding the χ^2 Test," *Ann. Math. Stats.*, vol. 4, 1933, p. 285.

(510) SHEPPARD, W. F., "The Fit of a Formula for Discrepant Observations," *Phil. Trans.*, Series A, vol. 228, 1927, p. 115.

(511) YULE, G. UDNY, "On the Application of the χ^2 Method to Association and Contingency Tables, with Experimental Illustrations," *Jour. Roy. Stat. Soc.*, vol. 85, 1922, p. 95.

CHAPTER 23. Sampling of Variables—Small Samples.

(Including some references to the theory of statistical inference.)

(512) BAKER, GEORGE A., "The Significance of the Product-moment Coefficient, with special reference to the Marginal Distributions," *Jour. Amer. Stat. Assoc.*, vol. 25, 1930, p. 387; and the related Paper: PEARSON, EGON S., "The Test of the Significance for the Correlation Coefficient," *Jour. Amer. Stat. Assoc.*, vol. 26, 1931, p. 128.

(513) BAKER, GEORGE A., "The Relation between the Means and Variances, Means Squared and Variances in Samples from Combinations of Normal Populations," *Ann. Math. Stats.*, vol. 2, 1931, p. 333.

(514) BAYES, T., "An Essay towards Solving a Problem in the Doctrine of Chances," *Phil. Trans.*, vol. 53, 1763, p. 370.

(515) BERKSON, JOSEPH, "Bayes' Theorem," *Ann. Math. Stats.*, vol. 1, 1930, p. 42.

(516) BOWLEY, A. L., "F. Y. Edgeworth's Contributions to Mathematical Statistics," published by the *Royal Statistical Society*, 1928.

(517) CAMP, BURTON H., "A New Generalisation of Tchebycheff's Statistical Inequality," *Bull. Amer. Math. Soc.*, vol. 28, 1922.

(518) CAMP, BURTON H., "Problems in Sampling," *Jour. Amer. Stat. Assoc.*, vol. 18, 1923, p. 964.

(519) CHESHIRE, L., B. OLDIS, and E. S. PEARSON, "Further Experiments on the Sampling Distribution of the Correlation Coefficient," *Jour. Amer. Stat. Assoc.*, vol. 27, 1932, p. 121.

(520) CHURCH, A. E. R., "On the Moments of the Distributions of Squared Standard Deviations for Samples of *N* drawn from an Indefinitely Large Population," *Biometrika*, vol. 17, 1925, p. 79.

(521) CHURCH, A. E. R., "On the Means and Squared Standard Deviations of Small Samples from any Population," *Biometrika*, vol. 18, 1926, p. 321.

(522) CRAIG, C. C., "Sampling when the Parent Population is of Pearson's Type III," *Biometrika*, vol. 21, 1929, p. 287.

(523) DODD, E. L., "The Convergence of General Means and the Invariance of Form of certain Frequency Functions," *Amer. Jour. Math.*, vol. 49, 1927.

(524) DODD, E. L., "The Greatest and the Least Variate under General Laws of Error," *Trans. Amer. Math. Soc.*, vol. 25, 1923, p. 525.

(525) DODD, E. L., "The Convergence of a General Mean of Measurements to the True Value," *Bull. Amer. Math. Soc.*, vol. 32, 1926.

(526) EZEKIEL, MORDECAI, "The Sampling Variability of Linear and Curvilinear Regression," *Ann. Math. Stats.*, vol. 1, 1930, p. 275.

(527) FISHER, R. A., "Inverse Probability," *Proc. Camb. Phil. Soc.*, vol. 26, 1930, p. 528.

(528) FISHER, R. A., "Inverse Probability and the Use of Likelihood," *Proc. Camb. Phil. Soc.*, vol. 28, 1932, p. 257.

(529) FISHER, R. A., "On the Probable Error of a Coefficient of Correlation deduced from a Small Sample," *Metron*, vol. 1, No. 4, 1921, p. 3. (See also refs. (434) and (435).)

(530) FISHER, R. A., "The General Sampling Distribution of the Multiple Correlation Coefficient," *Proc. Roy. Soc.*, A, vol. 121, 1928, p. 654.

(531) FISHER, R. A., "Moments and Product-moments of Sampling Distributions," *Proc. Lond. Math. Soc.*, vol. 30, 1928, p. 199.

(532) FISHER, R. A., and L. H. C. TIPPETT, "Limiting Forms of the Frequency-distribution of the Largest or Smallest Member of a Sample," *Proc. Camb. Phil. Soc.*, vol. 24, 1928, p. 180.

(533) FISHER, R. A., "On the Mathematical Foundations of Theoretical Statistics," *Phil. Trans.*, A, vol. 222, 1922, p. 309.

(534) FISHER, R. A., "The Theory of Statistical Estimation," *Proc. Camb. Phil. Soc.*, vol. 22, 1925, p. 700.

(535) FISHER, R. A., "On a Distribution Yielding the Error Functions of Several Well-known Statistics," *Proc. International Math. Congress at Toronto*, 1924, p. 805.

(536) FISHER, R. A., "Applications of 'Student's' Distribution" (and following tables by "Student"), *Metron*, vol. 5, No. 3, 1925, p. 90.

(537) GREENWOOD, M., and L. ISSERLIS, "An Historical Note on the Problem of Small Samples," *Jour. Roy. Stat. Soc.*, vol. 90, 1927, p. 347.

(538) HALL, PHILIP, "The Distribution of Means for Samples of Size *N* drawn from a Population in which the Variate takes Values between 0 and 1, all such Values being Equally Probable," *Biometrika*, vol. 19, 1927, p. 240.

(539) HOTELLING, H., "The Generalisation of 'Student's' Ratio," *Ann. Math. Stats.*, vol. 2, 1931, p. 360.

(540) HOTELLING, H., and MARGARET PABST, "Rank Correlation and Tests of Significance involving No Assumption of Normality," *Ann. Math. Stats.*, vol. 7, 1936, p. 29.

(541) IRWIN, J. O., "Mathematical Theorems involved in the Analysis of Variance," *Jour. Roy. Stat. Soc.*, vol. 94, 1931, p. 284.

(542) IRWIN, J. O., "On the Frequency-distribution of the Means of Samples from a Population having Any Law of Frequency with Finite Moments, etc.," *Biometrika*, vol. 19, 1927, p. 225, and vol. 21, 1929, p. 431.

(543) IRWIN, J. O., "On the Frequency-distribution of Any Number of Deviates from the Mean of a Sample from a Normal Population and the Partial Correlations between them," *Jour. Roy. Stat. Soc.*, vol. 92, 1929, p. 580.

(544) ISSERLIS, L., "On the Value of a Mean as calculated from a Sample," *Jour. Roy. Stat. Soc.*, vol. 81, 1918, p. 75.

(545) LE ROUX, J. M., "A Study of the Distribution of Variance in Small Samples," *Biometrika*, vol. 23, 1931, pp. 134–190.

(546) MEIDELL, H. BIRGER, "Sur un problème du calcul des probabilités et les statistiques mathématiques," *Comptes rendus*, vol. 175, 1922, p. 806.

(547) MOLINA, E. C., "Bayes' Theorem: An Expository Presentation," *Ann. Math. Stats.*, vol. 2, 1931, p. 23.

(548) NEYMAN, J., "Contributions to the Theory of Small Samples drawn from a Finite Population," *Revue Mensuelle de Statistique*, Office Central de Stat. de la République Polonaise, vol. 6, p. 1; reproduced in *Biometrika*, vol. 17, 1925, p. 472.

(549) NEYMAN, J., and E. S. PEARSON, "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," *Biometrika*, vol. 20A, 1928 and 1929, pp. 175 and 263.

(550) NEYMAN, J., and E. S. PEARSON, "On the Problem of *k* Samples," *Bull. de l'Acad. polonaise des Sci. et des Lettres*, Series A, 1931, p. 460.

(551) NEYMAN, J., and E. S. PEARSON, "On the Testing of Statistical Hypotheses in relation to Probability *a priori*," *Proc. Camb. Phil. Soc.*, vol. 29, 1933, p. 492.

(552) PEARSON, EGON S., and N. K. ADYANTHAYA, "The Distribution of Frequency Constants in Small Samples from Non-normal Symmetrical and Skew Populations," Preliminary Notice, *Biometrika*, vol. 20A, 1928, p. 356, and Second Paper, "Distribution of 'Student's' *z*," *Biometrika*, vol. 21, 1929, p. 259.

(553) PEARSON, EGON S., "Some Notes on Sampling Tests with Two Variables," *Biometrika*, vol. 21, 1929, p. 337.

(554) PEARSON, E. S., "The Test of Significance for the Correlation Coefficient," *Jour. Amer. Stat. Assoc.*, vol. 26, 1931, p. 128.

(555) PEARSON, EGON S., and J. NEYMAN, "On the Problem of Two Samples," *Bull. de l'Acad. polonaise des Sci. et des Lettres*, Series A, 1930, p. 73.

(556) PEARSON, E. S., "The Analysis of Variance in cases of Non-normal Variation," *Biometrika*, vol. 23, 1931, pp. 114 133.

(557) PEARSON, E. S., "The Test of Significance for the Correlation Coefficient—Some Further Results," *Jour. Amer. Stat. Assoc.*, vol. 27, 1932, p. 424.

(558) PEARSON, E. S., "Sampling Problems in Industry," *Jour. Roy. Stat. Soc.*, Suppl., vol. 1, 1934, p. 107.

(559) PEARSON, KARL, "On the Distribution of the Standard Deviation in Small Samples," *Biometrika*, vol. 10, 1915, p. 522.

(560) PEARSON, KARL, "The Fundamental Problem of Practical Statistics," *Biometrika*, vol. 13, 1920, p. 1.

(561) PEARSON, KARL, "Further Contributions to the Theory of Small Samples," *Biometrika*, vol. 17, 1925, p. 176.

(562) PEARSON, KARL, "Another Historical Note on the Theory of Small Samples," *Biometrika*, vol. 19, 1927, p. 207.

(562a) PEARSON, KARL, G. B. JEFFERY and E. M. ELDERTON, "On the Distribution of the First Product-moment Coefficient in Small Samples drawn from an Indefinitely Large Normal Population," *Biometrika*, vol. 21, 1929, p. 164.

(563) PEARSON, KARL, "Some Properties of 'Student's' *z*," *Biometrika*, vol. 23, 1931, p. 1.

(564) PEARSON, KARL, and BRENDA STOESSIGER, "Tables of the Probability Integrals of Symmetrical Frequency-curves in the case of Lower Powers such as arise in the Theory of Small Samples," *Biometrika*, vol. 22, 1931, p. 253.

(565) RIDER, PAUL R., "On Small Samples from certain Non-normal Universes," *Ann. Math. Stats.*, vol. 2, 1931, p. 48.

(566) RIDER, PAUL R., "A Note on Small Sample Theory," *Jour. Amer. Stat. Assoc.*, vol. 26, 1931, p. 172.

(567) RIDER, PAUL R., "On the Distribution of the Ratio of Mean to Standard Deviation in Small Samples from Non-normal Universes," *Biometrika*, vol. 21, 1929, p. 124.

(567a) RIDER, PAUL R., "A Survey of the Theory of Small Samples," *Ann. Maths.*, Oct. 1930, p. 577.

(568) RIDER, PAUL R., "On the Distribution of the Correlation Coefficient in Small Samples," *Biometrika*, vol. 24, 1932, p. 382.

(569) RIETZ, H. L., "Comments on the Applications of the Recently Developed Theory of Small Samples," *Jour. Amer. Stat. Assoc.*, vol. 26, 1931, p. 150.

(570) ROMANOVSKY, V., "Sulla probabilità a posteriori," *Giorn. dell' Istituto Italiano degli Attuari*," anno 2, 1931.

(571) ROMANOVSKY, V., "On the Criteria that Two Given Samples belong to the Same Normal Population," *Metron*, vol. 7, 1928, part 3, p. 3.

(572) ROMANOVSKY, V., "On the Moments of Means of Functions of One and More Random Variables," *Metron*, vol. 8, part 1, 1929, p. 251.

(572a) SHEWHART, W. A., and F. W. WINTERS, "Small Samples –New Experimental Results," *Jour. Amer. Stat. Assoc.*, vol. 23, 1928, pp. 144–153.

(573) SHOHAT, J. (Jacques Chokhate), "Inequalities for Moments of Frequency Functions and for Various Statistical Constants," *Biometrika*, vol. 21, 1929, p. 361.

(574) SMITH, C. D., "On Generalised Tchebycheff Inequalities in Mathematical Statistics," *Amer. Jour. Math.*, vol. 52, No. 1, 1930.

(575) SNEDECOR, G. W., *Calculation and Interpretation of Analysis of Variance and Covariance*; Collegiate Press, Ames, Iowa, 1934.

(576) SOPER, H. E., "The General Sampling Distribution of the Multiple Correlation Coefficient," *Jour. Roy. Stat. Soc.*, vol. 92, 1929, p. 445.

(577) SOPER, H. E., and Others, "On the Distribution of the Correlation Coefficient in Small Samples," *Biometrika*, vol. 11, 1916–17, p. 328.

(578) "SOPHISTER," "Discussion of Small Samples from an Infinite Skew Universe," *Biometrika*, vol. 20A, 1928, pp. 389–423.

(579) "STUDENT," "On the Probable Error of a Mean," *Biometrika*, vol. 6, 1908, p. 1.

(580) "STUDENT," "On the Probable Error of a Correlation Coefficient," *Biometrika*, vol. 6, 1908, p. 302. (The problem of the probable error with small samples.)

(581) "STUDENT," "On the z-Test"; followed by KARL PEARSON, "Further Remarks on the z-Test," *Biometrika*, vol. 23, 1931, pp. 407–415.

(582) TSCHUPROW, A. A., "On the Asymptotic Frequency-distributions of the Arithmetic Means of *n* Correlated Observations for Very Great Values of *n*," *Jour. Roy. Stat. Soc.*, vol. 88, 1925, p. 91.

(583) WILKS, S. S., "Certain Generalisations in the Analysis of Variance," *Biometrika*, vol. 24, 1932, p. 471.

(584) WISHART, JOHN, "The Generalised Product-moment Distribution in Samples from a Normal Multivariate Population," *Biometrika*, vol. 20A, 1928, p. 32.

(585) WISHART, JOHN, "The Correlation between Product-moments of Any Order in Samples from a Normal Population," *Proc. Roy. Soc. Edin.*, vol. 49, 1929, p. 1.

(586) WOO, T. L., "Tables for ascertaining the Significance or Non-significance of Association Measured by the Correlation Ratio," *Biometrika*, vol. 21, 1929, p. 1.

CHAPTER 24. Interpolation and Graduation.

(587) "Interpolation and Allied Tables." Reprint from *Nautical Almanac* for 1937; His Majesty's Stationery Office, 1936.

(588) PEARSON, KARL, *Tracts for Computers, II and III. On the Construction of Tables and on Interpolation*; Cambridge University Press, 1920.

(589) STEFFENSEN, J. F., *Some Recent Researches in the Theory of Statistics and Actuarial Science*; Cambridge: published for the *Institute of Actuaries* by the University Press, 1930.

(590) STEFFENSEN, J. F., *Interpolation*; Williams & Wilkins Co., Baltimore, 1927.

(591) WHITTAKER and ROBINSON, *The Calculus of Observations*; Blackie & Son, London; 2nd Ed., 1932.

The student who wishes to proceed further with the subject will probably find the last work cited the best for general use: it includes, of course, much besides interpolation. But (590) is very valuable for the advanced worker. All students are recommended to read the second lecture in the small work given under (589).

One can hardly give specific references, but the student will find much that is useful in the official publications of our own and other countries dealing with the construction of life-tables.

TABLES.

A. Tables Useful in Calculation.

(592) BARLOW'S *Tables of Squares, Cubes, Square-roots, Cube-roots and Reciprocals of all Integer Numbers up to 10,000*; E. & F. N. Spon, London and New York; new edition, 1930.

(593) COTSWORTH, M. B., *The Direct Calculator*, Series O. (Product table to 1000 × 1000.) M'Corquodale & Co., London.

(594) CRELLE, A. L., *Rechentafeln.* (Multiplication table giving all products up to 1000 × 1000.) Can be obtained with explanatory introduction in German or in English. G. Reimer, Berlin.

(595) ELDERTON, W. P. "Tables of Powers of Natural Numbers, and of the Sums of Powers of the Natural Numbers from 1 to 100" (gives powers up to seventh), *Biometrika,* vol. 2, p. 474—reproduced in (598).

(596) PETERS, J., *Neue Rechentafeln für Multiplikation und Division.* (Gives products up to 100 × 10,000: more convenient than Crelle for forming four-figure products. Introduction in English, French or German.) G. Reimer, Berlin.

(597) ZIMMERMANN, H., *Rechentafel,* nebst Sammlung häufig gebrauchter Zahlenwerthe. (Products of all numbers up to 100 × 1000: subsidiary tables of squares, cubes, square-roots, cube-roots and reciprocals, etc. for all numbers up to 1000 at the foot of the page.) W. Ernst & Son, Berlin; English edition, Asher & Co., London.

A number of useful tables will be found in the series "Tracts for Computers," published by the Cambridge University Press for the Department of Applied Statistics, University College, London. A list is usually given in the advertisement pages of the current issue of *Biometrika.*

B. Tables Useful in Statistical Work.

The more advanced student will probably find it indispensable to possess—

(598) *Tables for Statisticians and Biometricians, Part I* (edited by Karl Pearson), price 15s., from the *Biometrika* Office, University College, London, W.C. 1.

(599) Part II, price 30s., obtainable from the same address, contains tables of a more advanced character.

The following tables also contain much that is useful for modern statistical work :—

(600) *Tables of the Complete and Incomplete β-Function* (edited by Karl Pearson), price 55s.

(601) *Tables of the Incomplete Γ-Function* (edited by Karl Pearson), price 42s.

(602) *Tables of the Complete and Incomplete Elliptic Integrals,* price 12s. 6d.

The above are obtainable from the *Biometrika* Office, University College, London, W.C. 1.

(603) *Tracts for Computers, No. 1, Tables of the Digamma and Trigamma Functions,* price 3s.

(604) *Tracts for Computers, Nos. 4, 8 and 9, Logarithms of the Complete Γ-Function.*

(605) *Tracts for Computers, No. 15, Random Sampling Numbers,* by L. H. C. Tippett, price 3s. 9d.

(606) *British Association Mathematical Tables,* vol. 1, London, 1931; Office of the British Association, Burlington House, London, W. 1, price 10s., post free. (Circular and Hyperbolic Functions; Exponential Sine and Cosine Integrals; Factorial (Gamma) and Derived Functions; Integrals of Probability Integral.)

(607) *British Association Mathematical Tables,* vol. 6, London, 1936, price 40s. Bessel Functions, Part 1, Functions of Order 0 and 1.

(608) *Tables of the Higher Mathematical Functions* (edited by H. T. Davis), *Principia* Press, Bloomington, Indiana. (London: Williams & Norgate). *Part 1,* price 25s. (Historical Introduction, Tables of Γ- and Digamma-Functions.)

(609) *Part 2,* price 25s. (Tables of the Trigamma, Tetragamma, Pentagamma and Hexagamma Functions, of Bernoulli and Euler Numbers, of certain numbers facilitating the fitting of a polynomial.)

(610) KELLEY, T. L., "Tables to facilitate the Calculation of Partial Coefficients of Correlation and Regression Equations," *Bulletin of the University of Texas,* No. 27, 1916. (Tables giving the values of $1/\sqrt{(1-r_{13}^2)(1-r_{23}^2)}$ and $r_{13}r_{23}/\sqrt{(1-r_{13}^2)(1-r_{23}^2)}$.)

(611) MINER, J. R., *Tables of* $\sqrt{1-r^2}$ *and* $1-r^2$ *for use in Partial Correlation, etc.*; The Johns Hopkins Press, Baltimore, 1922. (Six-figure tables.)

(612) SALVOSA, L. R., "Tables of Pearson's Type III Function," *Ann. Math. Stats.,* vol. 1, 1930, p. 191.

References to Italian Literature.

In some respects the methods developed by the active school of Italian writers have diverged a good deal from those of English and American writers. The following bibliography, prepared by the kindness of Dr Silvio Orlandi, Manager of *Metron*, will serve as a guide to the student who wishes to broaden his outlook by making himself acquainted with such methods.

Books.

(613) BENINI, R., *Principi di statistica metodologica*; Unione Tipografica Editrice Torinese, Torino, 1926.

(614) BOLDRINI, M., *Statistica—Appunti per gli studenti*, voll. 2; Giuffrè, Milano, 1934–35.

(615) GINI, C., *Appunti di statistica metodologica*; Libreria Castellani, Roma, 1930–31. Traduzione spagnola: "Curso de Estadistica" (con un apendice matematico por L. Galvani), *Enciclopedia de Ciencias Yuridicas y Sociales*. Editorial Labour S.A., Barcelona, 1935.

(616) LIVI, L., *Elementi di statistica*; "Cedam," Padova, 1929.

(617) MORTARA, G., "Lezioni di statistica metodologica," Edite dal *Giornale degli Economisti e Rivista di Statistica*, Città di Castello, 1922.

(618) NICEFORO, A., *Il metodo statistico*; Messina. French translation, *La Méthode statistique*; Marcel Giard, Paris, 1925.

(619) PIETRA, G., *Statistica*, voll. 1 e 2; Giuffrè, Milano, 1934.

    See also

(620) *Trattato Elementare di Statistica*, diretto da C. Gini; Giuffrè, Milano, 1936. Vol. I, *Statistica Metodologica*; Vol. II, *Demografia*; Vol. III, *Antropometria e Biometria*; Vol. IV, *Statistica Economica*; Vol. V, *Statistica Economica*; Vol. VI, *Statistica sociale*.

General.

(621) GINI, C., "The Contributions of Italy to Modern Statistical Methods," *Journal of the Royal Statistical Society*, London, 1926.

(622) GINI, C., "Present Conditions and Future Progress of Statistics," *Journal of the American Statistical Association*, 1930.

Graphical Representation.

(623) GINI, C., "Sull' utilità delle rappresentazioni grafiche," *Giornale degli Economisti e Rivista di Statistica*, 1914.

(624) GINI, C., "Two Remarks on Graphs," *The Indian Journal of Statistics*, vol. 1, August 1934.

Interpolation and Extrapolation.

(625) CANTELLI, F. P., *Sull' adattamento di curve ad una serie di misure o di osservazioni*, Roma, 1905.

(626) GINI, C., "Considerazioni sull' interpolazione e la perequazione delle serie statistiche," *Metron*, vol. 1, fasc. 1, 1921.

(627) GINI, C., "Sull' interpolazione di una retta quando i valori della variabile indipendente sono affetti da errori accidentali," *Metron*, vol. 1, fasc. 4, 1921.

(628) GINI, C., "Ricerche sperimentali nel campo della interpolazione di serie statistiche," *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti*, 1923.

(629) MOGNO, R., "Di un metodo di interpolazione statistica," *Metron*, vol. 12, fasc. 2, 1934.

(630) PIETRA, G., "Interpolating Plane Curves," *Metron*, vol. 3, fasc. 3–4, 1924.

(631) PIETRA, G., "Dell' interpolazione parabolica nel caso in cui entrambi i valori delle variabili sono affetti da errori accidentali, *Metron*, vol. 9, fasc. 3–4, 1932.

(632) SALVEMINI, T., "Ricerche sperimentali sull' interpolazione grafica di istogrammi," *Metron*, vol. 11, fasc. 4, 1934.

(633) TEDESCHI, B., "Nuovo contributo al problema della interpolazione lineare," *Giornale dell' Istituto Italiano degli Attuari*, vol. 5, n. 2-3, 1934.

(634) VERONESE, G., *Contributo alle ricerche sperimentali nel campo dell' interpolazione statistica*; Padova.

Means, etc.

(635) GALVANI, L., "Sulla determinazione del centro di gravità e del centro mediano di una popolazione, con applicazione alla popolazione italiana censita al 1° dicembre 1921," *Metron*, vol. 11, n. 3, 1933.

(636) GINI, C., and L. GALVANI, "Di talune estensioni del concetto di media ai caratteri qualitativi," *Metron*, vol. 8, n. 1-2.

(637) GINI, C., M. BOLDRINI and A. VENERE, "Sui centri della popolazione e sulle loro applicazioni," *Metron*, vol. 11, n. 2.

Frequency and Probability.

(638) CANTELLI, F. P., "Sulla legge dei grandi numeri," *Memorie della R. Accad. dei Lincei*, 1916.

(639) CANTELLI, F. P., "Sulla probabilità come limite della frequenza," *Rendiconti della R. Accad. dei Lincei*, 1917.

(640) GINI, C., "Che cos'è la probabilità," *Rivista di Scienza*, 1908.

(641) GINI, C., "Il sesso dal punto di vista statistico," Cap. IV, pagg. 76-120, 125-131, *Istituto di Statistica della R. Università di Roma*.

(642) GINI, C., "Considerazioni sulle probabilità a posteriori e applicazione al rapporto dei sessi nelle nascite umane," *Studi Economico-Giuridici della R. Università di Cagliari*, 1911.

Variation and Concentration—"Transvariazione."

(643) CANTELLI, F. P., "Sulla differenza media con ripetizione," *Giornale degli Economisti e Rivista di Statistica*, February 1913.

(644) CASTELLANO, V., "Sulle relazioni fra curve di frequenza e curve di concentrazione e sui rapporti di concentrazione corrispondenti a determinate distribuzioni," *Metron*, vol. 10, n. 4, 1933.

(645) CASTELLANO, V., "Sugli indici relativi di variabilità e sulla concentrazione dei caratteri con segno," *Metron*, vol. 13, n. 1.

(646) DE FINETTI, B., "Sui metodi proposti per il calcolo della differenza media," *Metron*, vol. 9, n. 1, 1931.

(647) DE FINETTI, B., and PACIELLO, U., "Calcolo della differenza media," *Metron*, vol. 8, n. 3, 1930.

(648) DE VERGOTTINI, M., *Relazioni fra gli indici di variabilità dei fenomeni collettivi composti e quelli dei fenomeni collettivi semplici*; Failli, Roma, 1936.

(649) GALVANI, L., "Contributi alla determinazione degli indici di variabilità per alcuni tipi di distribuzione," *Metron*, vol. 9, n. 1, 1931.

(650) GALVANI, L., "Sulle curve di concentrazione relative a caratteri limitati e non limitati," *Metron*, vol. 10, n. 3, 1932.

(651) GINI, C., "Variabilità e Mutabilità, contributo allo studio delle distribuzioni e relazioni statistiche," *Studi Economico-Giuridici della R. Università di Cagliari*, 1912.

(652) GINI, C., "Indici di concentrazione e di dipendenza," *Biblioteca dell' Economista*, 5ª serie, 1910.

(653) GINI, C., "Sulla misura della concentrazione e della variabilità dei caratteri," *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti*, 1914.

(654) GINI, C., "Il concetto di transvariazione e le sue prime applicazioni," *Giornale degli Economisti e Rivista di Statistica*, 1916.

(655) GINI, C., "Di una estensione del concetto di scostamento medio e di alcune applicazioni alla misura della variabilità di caratteri qualitativi," *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti*, 1918.

(656) GINI, C., "Sul massimo degli indici di variabilità assoluta e sulle sue applicazioni agli indici di variabilità relativa e al rapporto di concentrazione," *Metron*, vol. 8, n. 3, 1930.

(657) GINI, C., "Intorno alle curve di concentrazione," *Metron*, vol. 9, n. 3-4, 1932.

(658) GINI, C., "Sull' influenza che il raggruppamento delle singole modalità esercita sul valore di alcuni indici statistici nel caso di serie sconnesse," *Metron,* vol. 12, n. 4, 1936.

(659) PIETRA, G., *Appunti intorno alla misura della variabilità e della concentrazione dei caratteri;* Bertero, Roma, 1915.

(660) PIETRA, G., "Delle relazioni tra gli indici di variabilità," *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti,* 1914–15, Parti I e II.

(661) PIETRA, G., "Intorno alla discordanza tra gli indici di variabilità e di concentrazione," *XXII Sessione dell' Istituto Internazionale di Statistica, Londra,* 1934.

(662) SAVÓRGNAN, F., "Intorno all' approssimazione di alcuni indici della distribuzione dei redditi," *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti,* 1915.

(663) VINCI, F., "Sui coefficienti di variabilità," *Metron,* vol. 1, n. 1, 1920.

Index-numbers and Other Statistical Measures.

(664) GINI, C., "Intorno al metodo dei residui dello Stuart Mill," *Studi Economico-Giuridici della R. Università di Cagliari,* 1910.

(665) GINI, C., "Quelques considérations au sujet de la construction des nombres indices des prix et des questions analogues. Contribution à l'étude des méthodes d'élimination," *Metron,* vol. 3, n. 1, 1924.

(666) GINI, C., "On the Circular Test of Index-numbers," *Metron,* vol. 9, n. 2. 1931.

(667) GINI, C., "Tavole di mortalità della popolazione italiana" (in collaborazione con L. Galvani), *Annali di Statistica,* Serie 6, vol. 8, 1931.

(668) GINI, C., "Sur une méthode pour determiner le nombre moyen des enfants légitimes par mariages," *Revue de l'Institut International de Statistique,* 1934.

(669) GINI, C., "Sur la mesure de la fécondité des mariages," *Bulletin de l'Institut International de Statistique,* 1934.

(670) GINI, C., "On a Method for Calculating the Infantile Death-rate according to the Month of Death," *Revue de l'Institut International de Statistique,* 1934.

(671) GINI, C., "Su la determinazione dei quozienti di eliminazione e in particolare sui metodi delle durate esatte e delle durate medie nella ipotesi di saggi istantanei di eliminazione costanti," *Metron,* vol. 12, n. 3, 1935.

(672) GINI, C., "Methods of Eliminating the Influence of Several Groups of Factors," *Econometrica,* January 1937.

Statistical Relations.

(673) GINI, C., "Di una misura della dissomiglianza tra due gruppi di quantità e delle sue applicazioni allo studio delle relazioni statistiche," *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti,* 1914.

(674) GINI, C., "Nuovi contributi alla teoria delle relazioni statistiche," *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti,* 1915.

(675) GINI, C., "Indici di omofilia e di rassomiglianza e loro relazioni col coefficiente di correlazione e con gli indici di attrazione," *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti,* 1915.

(676) GINI, C., "Sul criterio di concordanza tra due caratteri," *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti,* 1916.

(677) GINI, C., "Indici di concordanza," *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti,* 1916.

(678) GINI, C., "Sulle relazioni tra le intensità cograduate di due caratteri," *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti,* 1917.

(679) GINI, C., "Sull' influenza che il raggruppamento delle singole modalità esercita sul valore di alcuni indici statistici nel caso di serie sconnesse," *Metron,* vol. 12, n. 4, 1936.

(680) PIETRA, G., "The Theory of Statistical Relations, with Special Reference to Cyclical Series," *Metron,* vol. 4, n. 3–4, 1925.

APPENDIX TABLES.

.

APPENDIX TABLE 1.

Normal Curve. *Ordinates of the Normal Curve of Errors of Unit Area at every Tenth of the Standard Deviation, with First and Second Differences. The value of the central ordinate at zero is* $1/\sqrt{2\pi}$.

x/σ.	y.	$\Delta^1(-)$.	Δ^2.	x/σ.	y.	$\Delta^1(-)$.	Δ^2.
0·0	0·39894	199	− 392	2·5	0·01753	395	+ 79
0·1	·39695	591	− 374	2·6	·01358	316	+ 66
0·2	·39104	965	− 347	2·7	·01042	250	+ 53
0·3	·38139	1312	− 308	2·8	·00792	197	+ 45
0·4	·36827	1620	− 265	2·9	·00595	152	+ 36
0·5	·35207	1885	− 212	3·0	·00443	116	+ 27
0·6	·33322	2097	− 159	3·1	·00327	89	+ 23
0·7	·31225	2256	− 104	3·2	·00238	66	+ 17
0·8	·28969	2360	− 52	3·3	·00172	49	+ 13
0·9	·26609	2412	0	3·4	·00123	36	+ 10
1·0	·24197	2412	+ 46	3·5	·00087	26	+ 7
1·1	·21785	2366	+ 84	3·6	·00061	19	+ 6
1·2	·19419	2282	+118	3·7·	·00042	13	+ 4
1·3	·17137	2164	+143	3·8	·00029	9	+ 2
1·4	·14973	2021	+161	3·9	·00020	7	+ 3
1·5	·12952	1860	+173	4·0	·00013	4	—
1·6	·11092	1687	+177	4·1	·00009	3	—
1·7	·09405	1510	+177	4·2	·00006	2	—
1·8	·07895	1333	+170	4·3	·00004	2	—
1·9	·06562	1163	+162	4·4	·00002	—	—
2·0	·05399	1001	·+150	4·5	·00002	—	—
2·1	·04398	851	+137	4·6	·00001	—	—
2·2	·03547	714	+120	4·7	·00001	—	—
2·3	·02833	594	+108	4·8	·00000	—	—
2·4	·02239	486	+ 91				

Precision of Interpolation.—Owing to the magnitude of the second differences, simple interpolation near the beginning of the table may give an error up to 5 in the fourth place; the use of second differences will bring this down to 1 or 2 in the last place, third differences being small. Where third differences are greatest, in the neighbourhood of $x/\sigma=0\cdot6$, the error may be as large as 3 in the last place unless the third difference is used.

APPENDIX TABLE 2.

Normal Curve. *The Proportion, A, of the Whole Area of the Normal Curve lying to the Left of the Ordinate at Deviation x/σ, tabulated at every Tenth of the Standard Deviation, with First and Second Differences.*

x/σ.	A.	$\Delta^1(+)$.	$\Delta^2(-)$.	x/σ.	A.	$\Delta^1(+)$.	$\Delta^2(-)$.
0·0	0·50000	3983	40	2·5	0·99379	155	36
0·1	·53983	3943	78	2·6	·99534	119	28
0·2	·57926	3865	114	2·7	·99653	91	22
0·3	·61791	3751	147	2·8	·99744	69	17
0·4	·65542	3604	175	2·9	·99813	52	14
0·5	·69146	3429	200	3·0	·99865	38	10
0·6	·72575	3229	219	3·1	·99903	28	7
0·7	·75804	3010	230	3·2	·99931	21	7
0·8	·78814	2780	240	3·3	·99952	14	3
0·9	·81594	2540	241	3·4	·99966	11	4
1·0	·84134	2299	239	3·5	·99977	7	—
1·1	·86433	2060	233	3·6	·99984	5	—
1·2	·88493	1827	223	3·7	·99989	4	—
1·3	·90320	1604	209	3·8	·99993	2	—
1·4	·91924	1395	194	3·9	·99995	2	—
1·5	·93319	1201	178	4·0	·99997	1	—
1·6	·94520	1023	159	4·1	·99998	1	—
1·7	·95543	864	143	4·2	·99999	—	·
1·8	·96407	721	124	4·3	·99999	—	—
1·9	·97128	597	108	4·4	·99999	—	—
2·0	·97725	489	93				
2·1	·98214	396	78				
2·2	·98610	318	66				
2·3	·98928	252	53				
2·4	·99180	199	44				

A attains the exact value 0·99999 between 4·26 and 4·27.

Precision of Interpolation.—Simple interpolation may lead to an error of 3 or 4 at most in the fourth place of decimals in the region where second differences are large; the use of the second difference will bring this down to 2 or 3 in the last place, the largest errors tending to occur at the beginning of the table, where the third difference may be used if the greatest possible precision is desired.

APPENDIX TABLE 3.

Normal Curve. *The Probability, P, of an Observation lying Outside the Limits* $\pm x/\sigma$ *in the Normal Curve of Errors.* $P = 2(1 - A)$, *where A is the area given by the preceding table.*

x/σ.	P.	$\Delta^1(-)$.	$\Delta^2(+)$.	x/σ.	P.	$\Delta^1(-)$.	$\Delta^2(+)$.
0·0	1·00000	7966	80	2·5	·01242	310	71
0·1	0·92034	7886	·156	2·6	·00932	239	57
0·2	·84148	7730	228	2·7	·00693	182	44
0·3	·76418	7502	294	2·8	·00511	138	35
0·4	·68916	7208	351	2·9	·00373	103	27
0·5	·61708	6857	399	3·0	·00270	76	19
0·6	·54851	6458	436	3·1	·00194	57	17
0·7	·48393	6022	463	3·2	·00137	40	10
0·8	·42371	5559	478	3·3	·00097	30	10
0·9	·36812	5081	483	3·4	·00067	20	5
1·0	·31731	4598	479	3·5	·00047	15	—
1·1	·27133	4119	465	3·6	·00032	10	—
1·2	·23014	3654	445	3·7	·00022	8	—
1·3	·19360	3209	419	3·8	·00014	4	—
1·4	·16151	2790	389	3·9	·00010	4	—
1·5	·13361	2401	354	4·0	·00006	2	—
1·6	·10960	2047	320	4·1	·00004	1	—
1·7	·08913	1727	284	4·2	·00003	1	—
1·8	·07186	1443	250	4·3	·00002	1	—
1·9	·05743	1193	216	4·4	·00001	—	—
				4·5	·00001	—	—
2·0	·04550	977	185				
2·1	·03573	792	156				
2·2	·02781	636	131				
2·3	·02145	505	107				
2·4	·01640	398	88				

P attains the exact value 0·00001 between 4·41 and 4·42.

Precision of Interpolation.—Simple interpolation may lead to errors of 5 or 6 in the fourth place of decimals, where second differences are large. Using second differences as well, the error will not exceed about 5 in the last place, near the beginning of the table, where the third difference may be brought in if desired.

APPENDIX TABLE 4A.

*Values of the χ^2 Integral for One Degree of Freedom for Values of χ^2
from $\chi^2 = 0$ to $\chi^2 = 1$ by steps of 0·01.*

χ^2	P	Δ	χ^2	P	Δ
0	1·00000	7966	0·50	0·47950	436
0·01	0·92034	3280	0·51	0·47514	430
0·02	0·88754	2505	0·52	0·47084	423
0·03	0·86249	2101	0·53	0·46661	418
0·04	0·84148	1842	0·54	0·46243	411
0·05	0·82306	1656	0·55	0·45832	406
0·06	0·80650	1516	0·56	0·45426	400
0·07	0·79134	1404	0 57	0·45026	395
0·08	0·77730	1312	0·58	0·44631	389
0·09	0·76418	1235	0·59	0·44242	384
0·10	0·75183	1169	0·60	0·43858	379
0·11	0·74014	1111	0·61	0·43479	374
0·12	0·72903	1060	0·62	0·43105	369
0·13	0·71843	1015	0·63	0·42736	365
0·14	0·70828	974	0·64	0·42371	360
0·15	0·69854	938	0·65	0·42011	355
0·16	0·68916	905	0·66	0·41656	351
0·17	0·68011	874	0·67	0·41305	346
0·18	0·67137	845	0·68	0·40959	343
0·19	0·66292	820	0·69	0·40616	338
0·20	0·65472	795	0·70	0·40278	334
0·21	0·64677	773	0·71	0·39944	330
0·22	0·63904	752	0·72	0·39614	326
0·23	0·63152	731	0·73	0·39288	322
0·24	0 62421	713	0·74	0·38966	318
0·25	0·61708	696	0·75	0·38648	315
0·26	0·61012	679	0·76	0·38333	311
0·27	0·60333	663	0·77	0·38022	308
0·28	0·59670	648	0·78	0·37714	304
0·29	0·59022	634	0·79	0·37410	301
0·30	0 58388	620	0·80	0·37109	297
0·31	0·57768	607	0·81	0·36812	294
0·32	0 57161	595	0·82	0·36518	291
0·33	0·56566	583	0·83	0·36227	287
0·34	0·55983	572	0·84	0·35940	285
0·35	0·55411	560	0·85	0·35655	281
0·36	0·54851	551	0·86	0·35374	278
0 37	0·54300	540	0·87	0·35096	276
0·38	0·53760	530	0·88	0·34820	272
0·39	0·53230	521	0·89	0·34548	270
0·40	0·52709	512	0·90	0·34278	267
0·41	0·52197	503	0·91	0·34011	264
0·42	0·51694	495	0·92	0·33747	261
0·43	0·51199	487	0·93	0·33486	258
0·44	0·50712	479	0·94	0·33228	256
0·45	0·50233	471	0·95	0·32972	253
0·46	0·49762	463	0·96	0·32719	251
0·47	0·49299	457	0·97	0·32468	248
0·48	0·48842	449	0·98	0 32220	246
0·49	0·48393	443	0·99	0·31974	243
0·50	0·47950	436	1·00	0·31731	241

APPENDIX TABLE 4B.

Values of the χ^2 Integral for One Degree of Freedom for Values of χ^2 from $\chi^2 = 1$ to $\chi^2 = 10$ by steps of 0.1.

χ^2	P	Δ	χ^2	P	Δ
1·0	0·31731	2304	5·5	0·01902	106
1·1	0·29427	2095	5·6	0·01796	99
1·2	0·27332	1911	5·7	0·01697	94
1·3	0·25421	1749	5·8	0·01603	89
1·4	0·23672	1605	5·9	0·01514	83
1·5	0·22067	1477	6·0	0·01431	79
1·6	0·20590	1361	6·1	0·01352	74
1·7	0·19229	1258	6·2	0·01278	71
1·8	0·17971	1163	6·3	0·01207	66
1·9	0·16808	1078	6·4	0·01141	62
2·0	0·15730	1000	6·5	0·01079	59
2·1	0·14730	929	6·6	0·01020	56
2·2	0·13801	864	6·7	0·00964	52
2·3	0·12937	803	6·8	0·00912	50
2·4	0·12134	749	6·9	0·00862	47
2·5	0·11385	699	7·0	0·00815	44
2·6	0·10686	651	7·1	0·00771	42
2·7	0·10035	609	7·2	0·00729	39
2·8	0·09426	568	7·3	0·00690	38
2·9	0·08858	532	7·4	0·00652	35
3·0	0·08326	497	7·5	0·00617	33
3·1	0·07829	465	7·6	0·00584	32
3·2	0·07364	436	7·7	0·00552	30
3·3	0·06928	408	7·8	0·00522	28
3·4	0·06520	383	7·9	0·00491	26
3·5	0·06137	359	8·0	0·00468	25
3·6	0·05778	337	8·1	0·00443	24
3·7	0·05441	316	8·2	0·00419	23
3·8	0·05125	296	8·3	0·00396	21
3·9	0·04829	279	8·4	0·00375	20
4·0	0·04550	262	8·5	0·00355	19
4·1	0·04288	246	8·6	0·00336	18
4·2	0·04042	231	8·7	0·00318	17
4·3	0·03811	217	8·8	0·00301	16
4·4	0·03594	205	8·9	0·00285	15
4·5	0·03389	192	9·0	0·00270	14
4·6	0·03197	181	9·1	0·00256	14
4·7	0·03016	170	9·2	0·00242	13
4·8	0·02846	160	9·3	0·00229	12
4·9	0·02686	151	9·4	0·00217	12
5·0	0·02535	142	9·5	0·00205	10
5·1	0·02393	134	9·6	0·00195	11
5·2	0·02259	126	9·7	0·00184	10
5·3	0·02133	119	9·8	0·00174	9
5·4	0·02014	112	9·9	0·00165	8
5·5	0·01902	106	10·0	0·00157	8

t-Table. *The Proportion of the Area of the Curve* $y = \dfrac{y_0}{\left(1+\dfrac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}}$ *of Unit Area lying to*

0 to 6, and for values

(Condensed to three figures from the four-figure tables by "Student" in *Metron*, "Student," who has also very kindly supplied

t.	ν = 1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
0	0·500	0·500	0·500	0·500	0·500	0·500	0·500	0·500	0·500	0·500
0·1	·532	·535	·537	·537	·538	·538	·538	·539	·539	·539
·2	·563	·570	·573	·574	·575	·576	·576	·577	·577	·577
·3	·593	·604	·608	·610	·612	·613	·614	·614	·614⁵	·615
·4	·621	·636	·642	·645	·647	·648⁵	·649⁵	·650	·651	·651
·5	·648	·667	·674	·678	·681	·683	·684	·685	·685⁵	·686
·6	·672	·695	·705	·710	·713	·715	·716	·717	·718	·719
·7	·694	·722	·733	·739	·742	·745	·747	·748	·749	·750
·8	·715	·746	·759	·766	·770	·773	·775	·777	·778	·779
·9	·733	·768	·783	·790⁵	·795	·799	·801	·803	·804	·805
1·0	·750	·789	·804⁵	·813	·818	·822	·825	·827	·828	·830
1·1	·765	·807	·824	·833⁵	·839	·843	·846	·848	·850	·851
1·2	·779	·823⁵	·842	·852	·858	·862	·865	·868	·870	·871
1·3	·791	·838	·858	·868	·875	·879	·883	·885	·887	·889
1·4	·803	·852	·872	·883	·890	·894⁵	·898	·900⁵	·902⁵	·904
1·5	·813	·864	·885	·896	·903	·908	·911	·914	·916	·918
1·6	·822	·875	·896	·908	·915	·920	·923	·926	·928	·930
1·7	·831 ⁻	·884	·906	·918	·925	·930	·933⁵	·936	·938	·940
1·8	·839	·893	·915	·927	·934	·939	·943	·945	·947	·949
1·9	·846	·901	·923	·935	·942	·947	·950	·953	·955	·957
2·0	·852	·908	·930	·942	·949	·954	·957	·960	·962	·963
2·1	·858⁵	·915	·937	·948	·955	·960	·963	·965⁵	·967	·969
2·2	·864	·921	·942	·954	·960⁵	·965	·968	·970⁵	·972	·974
2·3	·869⁵	·926	·947⁵	·958⁵	·965	·969	·972⁵	·975	·976⁵	·978
2·4	·874	·931	·952	·963	·969	·973	·976	·978	·980	·981
2·5	·879	·935	·956	·967	·973	·977	·979⁵	·981⁵	·983	·984
2·6	·883	·939	·960	·970	·976	·980	·982	·984	·986	·987
2·7	·887	·943	·963	·973	·979	·982	·985	·986⁵	·988	·989
2·8	·891	·946	·966	·976	·981	·984	·987	·988	·990	·991
2·9	·894	·949	·969	·978	·983	·986	·988⁵	·990	·991	·992
3·0	·898	·952	·971	·980	·985	·988	·990	·991⁵	·992⁵	·993
3·1	·901	·955	·973	·982	·987	·989	·991	·993	·994	·994
3·2	·904	·957	·975	·983⁵	·988	·991	·992⁵	·994	·995	·995
3·3	·906	·960	·977	·985	·989	·992	·993	·995	·995	·996
3·4	·909	·962	·979	·986	·990	·993	·994	·995	·996	·997
3·5	·911	·964	·980	·988	·991	·994	·995	·996	·997	·997
3·6	·914	·965	·982	·989	·992	·994	·996	·996⁵	·997	·998
3·7	·916	·967	·983	·990	·993	·995	·996	·997	·997⁵	·998
3·8	·918	·969	·984	·990	·994	·995⁵	·997	·997	·998	·998
3·9	·920	·970	·985	·991	·994	·996	·997	·998	·998	·998⁵
4·0	·922	·971	·986	·992	·995	·996	·997	·998	·998	·999
4·1	·924	·973	·987	·993	·995	·997	·998	·998	·999	·999
4·2	·926	·974	·988	·993	·996	·997	·998	·998⁵	·999	·999
4·3	·927	·975	·988	·994	·996	·997⁵	·998	·999	·999	·999
4·4	·929	·976	·989	·994	·996⁵	·998	·998	·999	·999	·999
4·5	·930	·977	·990	·995	·997	·998	·999	·999	·999	·999
4·6	·932	·978	·990	·995	·997	·998	·999	·999	·999	·999⁵
4·7	·933	·979	·991	·995	·997	·998	·999	·999	·999	1·000
4·8	·935	·980	·991	·996	·998	·998⁵	·999	·999	·999⁵	
4·9	·936	·980	·992	·996	·998	·999	·999	·999	1·000	
5·0	·937	·981	·992	·996	·998	·999	·999	·999⁵		
5·1	·938	·982	·993	·996⁵	·998	·999	·999	·999⁵		
5·2	·939⁵	·982⁵	·993	·997	·998	·999	·999	1·000		
5·3	·941	·983	·993	·997	·998	·999	·999			
5·4	·942	·984	·994	·997	·998⁵	·999	·999⁵			
5·5	·943	·984	·994	·997	·999	·999	·999⁵			
5·6	·944	·985	·994	·997⁵	·999	·999	1·000			
5·7	·945	·985	·995	·998	·999	·999				
5·8	·946	·986	·995	·998	·999	·999				
5·9	·947	·986	·995	·998	·999	·999⁵				

TABLE 5.

the Left of the Ordinate of Deviation t, for values of t proceeding by intervals of 0·1 from

of v from 1 to 20.

vol. 5, 1925, and published by permission of the proprietors of *Metron* and
a few corrections to the original tables.)

t.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.
0	0·500	0·500	0·500	0·500	0·500	0·500	0·500	0·500	0·500	0·500
0·1	·539	·539	·539	·539	·539	·539	·539	·539	·539	·539
·2	·577	·578	·578	·578	·578	·578	·578	·578	·578	·578
·3	·615	·615	·615⁵	·616	·616	·616	·616	·616	·616	·616
·4	·652	·652	·652	·652	·653	·653	·653	·653	·653	·653
·5	·686⁵	·687	·687	·688	·688	·688	·688	·688	·689	·689
·6	·720	·720	·721	·721	·721	·721⁵	·722	·722	·722	·722
·7	·751	·751	·752	·752	·753	·753	·753	·754	·754	·754
·8	·780	·780	·781	·781⁵	·782	·782	·783	·783	·783	·783
·9	·806	·807	·808	·808	·809	·809	·810	·810	·810	·811
1·0	·831	·831⁵	·832	·833	·833	·834	·834	·835	·835	·835
1·1	·853	·853⁵	·854	·855	·856	·856	·857	·857	·857⁵	·858
1·2	·872	·873	·874	·875	·876	·876	·877	·877	·878	·878
1·3	·890	·891	·892	·893	·893	·894	·894⁵	·895	·895	·896
1·4	·905⁵	·907	·907⁵	·908	·909	·909	·910	·911	·911	·912
1·5	·919	·920	·921	·922	·923	·923⁵	·924	·924⁵	·925	·925
1·6	·931	·932	·933	·934	·935	·935	·936	·936⁵	·937	·937
1·7	·941	·943	·943⁵	·944	·945	·946	·946	·947	·947	·948·
1·8	·950	·951⁵	·952⁵	·953	·954	·955	·955	·956	·956	·956⁵
1·9	·958	·959	·960	·961	·962	·962	·963	·963	·964	·964
2·0	·965	·966	·967	·967	·968	·969	·969	·970	·970	·970
2·1	·970	·971	·972	·973	·973⁵	·974	·974⁵	·975	·975	·976
2·2	·975	·976	·977	·977	·978	·979	·979	·979	·980	·980
2·3	·979	·980	·981	·981	·982	·982	·983	·983	·983⁵	·984
2·4	·982	·983	·984	·985	·985	·985⁵	·986	·986	·987	·987
2·5	·985	·986	·987	·987	·988	·988	·988⁵	·989	·989	·989
2·6	·988	·988	·989	·989⁵	·990	·990	·991	·991	·991	·991
2·7	·990	·990	·991	·991	·992	·992	·992	·993	·993	·993
2·8	·991	·992	·992⁵	·993	·993	·994	·994	·994	·994	·994⁵
2·9	·993	·993	·994	·994	·994⁵	·994⁵	·995	·995	·995	·996
3·0	·994	·994⁵	·995	·995	·995⁵	·996	·996	·996	·996	·996⁵
3·1	·995	·995	·996	·996	·996	·997	·997	·997	·997	·997
3·2	·996	·996	·996⁵	·997	·997	·997	·997	·997⁵	·998	·998
3·3	·996⁵	·997	·997	·997	·998	·998	·998	·998	·998	·998
3·4	·997	·997	·998	·998	·998	·998	·998	·998	·998⁵	·999
3·5	·997⁵	·998	·998	·998	·998	·998⁵	·999	·999	·999	·999
3·6	·998	·998	·998	·999	·999	·999	·999	·999	·999	·999
3·7	·998	·998⁵	·999	·999	·999	·999	·999	·999	·999	·999
3·8	·998⁵	·999	·999	·999	·999	·999	·999	·999	·999	·999
3·9	·999	·999	·999	·999	·999	·999	·999	·999⁵	·999⁵	1·000
4·0	·999	·999	·999	·999	·999	·999⁵	·999⁵	1·000	1·000	
4·1	·999	·999	·999	·999⁵	·999⁵	1·000	1·000			
4·2	·999	·999	·999⁵	1·000	1·000					
4·3	·999	·999⁵	1·000							
4·4	·999⁵	1·000								
4·5	·999⁵									
4·6	1·000									

Note.—The methods by which "Student" calculated the *Metron* tables are
explained in notes by him and R. A. Fisher in that journal, vol. 5, part 3, 1925,
pp. 18–24. The four figures of those values have been rounded up to three in
the above table, except when the four-figure value concluded with a 5, in which
case it is shown in full. In columns in which values greater than 0·9995 occur
the first is written 1·000 and the remainder left blank.

APPENDIX TABLE 6A.

5 PER CENT. POINTS OF THE DISTRIBUTION OF z.

		Values of ν_1.									
		1.	2.	3.	4.	5.	6.	8.	12.	24.	∞.
	1	2·5421	2·6479	2·6870	2·7071	2·7194	2·7276	2·7380	2·7484	2·7588	2·7693
	2	1·4592	1·4722	1·4765	1·4787	1·4800	1·4808	1·4819	1·4830	1·4840	1·4851
	3	1·1577	1·1284	1·1137	1·1051	1·0994	1·0953	1·0899	1·0842	1·0781	1·0716
	4	1·0212	·9690	·9429	·9272	·9168	·9093	·8993	·8885	·8767	·8639
	5	·9441	·8777	·8441	·8236	·8097	·7997	·7862	·7714	·7550	·7368
	6	·8948	·8188	·7798	·7558	·7394	·7274	·7112	·6931	·6729	·6499
	7	·8606	·7777	·7347	·7080	·6896	·6761	·6576	·6369	·6134	·5862
	8	·8355	·7475	·7014	·6725	·6525	·6378	·6175	·5945	·5682	·5371
	9	·8163	·7242	·6757	·6450	·6238	·6080	·5862	·5613	·5324	·4979
	10	·8012	·7058	·6553	·6232	·6009	·5843	·5611	·5346	·5035	·4657
	11	·7889	·6909	·6387	·6055	·5822	·5648	·5406	·5126	·4795	·4387
	12	·7788	·6786	·6250	·5907	·5666	·5487	·5234	·4941	·4592	·4156
	13	·7703	·6682	·6134	·5783	·5535	·5350	·5089	·4785	·4419	·3957
	14	·7630	·6594	·6036	·5677	·5423	·5233	·4964	·4649	·4269	·3782
	15	·7568	·6518	·5950	·5585	·5326	·5131	·4855	·4532	·4138	·3628
	16	·7514	·6451	·5876	·5505	·5241	·5042	·4760	·4428	·4022	·3490
	17	·7466	·6393	·5811	·5434	·5166	·4964	·4676	·4337	·3919	·3366
	18	·7424	·6341	·5753	·5371	·5099	·4894	·4602	·4255	·3827	·3253
	19	·7386	·6295	·5701	·5315	·5040	·4832	·4535	·4182	·3743	·3151
Values of ν_2.	20	·7352	·6254	·5654	·5265	·4986	·4776	·4474	·4116	·3668	·3057
	21	·7322	·6216	·5612	·5219	·4938	·4725	·4420	·4055	·3599	·2971
	22	·7294	·6182	·5574	·5178	·4894	·4679	·4370	·4001	·3536	·2892
	23	·7269	·6151	·5540	·5140	·4854	·4636	·4325	·3950	·3478	·2818
	24	·7246	·6123	·5508	·5106	·4817	·4598	·4283	·3904	·3425	·2749
	25	·7225	·6097	·5478	·5074	·4783	·4562	·4244	·3862	·3376	·2685
	26	·7205	·6073	·5451	·5045	·4752	·4529	·4209	·3823	·3330	·2625
	27	·7187	·6051	·5427	·5017	·4723	·4499	·4176	·3786	·3287	·2569
	28	·7171	·6030	·5403	·4992	·4696	·4471	·4146	·3752	·3248	·2516
	29	·7155	·6011	·5382	·4969	·4671	·4444	·4117	·3720	·3211	·2466
	30	·7141	·5994	·5362	·4947	·4648	·4420	·4090	·3691	·3176	·2419
	60	·6933	·5738	·5073	·4632	·4311	·4064	·3702	·3255	·2654	·1644
	∞	·6729	·5486	·4787	·4319	·3974	·3706	·3309	·2804	·2085	0

APPENDIX TABLE 6B.

(Reproduced by kind permission of Prof. R. A. Fisher and Messrs Oliver & Boyd
from the former's "*Statistical Methods for Research Workers*.")

1 PER CENT. POINTS OF THE DISTRIBUTION OF z.

		Values of ν_1.									
		1.	2.	3.	4.	5.	6.	8.	12.	24.	∞.
	1	4·1535	4·2585	4·2974	4·3175	4·3297	4·3379	4·3482	4·3585	4·3689	4·3794
	2	2·2950	2·2976	2·2984	2·2988	2·2901	2·2992	2·2994	2·2997	2·2999	2·3001
	3	1·7649	1·7140	1·6915	1·6786	1·6703	1·6645	1·6569	1·6489	1·6404	1·6314
	4	1·5270	1·4452	1·4075	1·3856	1·3711	1·3609	1·3473	1·3327	1·3170	1·3000
	5	1·3943	1·2929	1·2449	1·2164	1·1974	1·1838	1·1656	1·1457	1·1239	1·0997
	6	1·3103	1·1955	1·1401	1·1068	1·0843	1·0680	1·0460	1·0218	·9948	·9643
	7	1·2526	1·1281	1·0672	1·0300	1·0048	·9864	·9614	·9335	·9020	·8658
	8	1·2106	1·0787	1·0135	·9734	·9459	·9259	·8983	·8673	·8319	·7904
	9	1·1786	1·0411	·9724	·9299	·9006	·8791	·8494	·8157	·7769	·7305
	10	1·1535	1·0114	·9399	·8954	·8646	·8419	·8104	·7744	·7324	·6816
	11	1·1333	·9874	·9136	·8674	·8354	·8116	·7785	·7405	·6958	·6408
	12	1·1166	·9677	·8919	·8443	·8111	·7864	·7520	·7122	·6649	·6061
	13	1·1027	·9511	·8737	·8248	·7907	·7652	·7295	·6882	·6386	·5761
	14	1·0909	·9370	·8581	·8082	·7732	·7471	·7103	·6675	·6159	·5500
15	15	1·0807	·9249	·8448	·7939	·7582	·7314	·6937	·6496	·5961	·5269
	16	1·0719	·9144	·8331	·7814	·7450	·7177	·6791	·6339	·5788	·5064
	17	1·0641	·9051	·8229	·7705	·7335	·7057	·6663	·6199	·5630	·4879
	18	1·0572	·8970	·8138	·7607	·7232	·6950	·6549	·6075	·5491	·4712
	19	1·0511	·8897	·8057	·7521	·7140	·6854	·6447	·5964	·5366	·4560
	20	1·0457	·8831	·7985	·7443	·7058	·6768	·6355	·5864	·5253	·4421
	21	1·0408	·8772	·7920	·7372	·6984	·6690	·6272	·5773	·5150	·4294
	22	1·0363	·8719	·7860	·7309	·6916	·6620	·6196	·5691	·5056	·4176
	23	1·0322	·8670	·7806	·7251	·6855	·6555	·6127	·5615	·4969	·4068
	24	1·0285	·8626	·7757	·7197	·6799	·6496	·6064	·5545	·4890	·3967
	25	1·0251	·8585	·7712	·7148	·6747	·6442	·6006	·5481	·4816	·3872
	26	1·0220	·8548	·7670	·7103	·6699	·6392	·5952	·5422	·4748	·3784
	27	1·0191	·8513	·7631	·7062	·6655	·6346	·5902	·5367	·4685	·3701
	28	1·0164	·8481	·7595	·7023	·6614	·6303	·5856	·5316	·4626	·3624
	29	1·0139	·8451	·7562	·6987	·6576	·6263	·5813	·5269	·4570	·3550
	30	1·0116	·8423	·7531	·6954	·6540	·6226	·5773	·5224	·4519	·3481
	60	·9784	·8025	·7086	·6472	·6028	·5687	·5189	·4574	·3746	·2352
	∞	·9462	·7636	·6651	·5999	·5522	·5152	·4604	·3908	·2913	0

Values of ν_2.

APPENDIX TABLE 6C.

(Reproduced by kind permission of Prof. R. A. Fisher, Dr W. E. Deming and Messrs Oliver & Boyd from Prof. Fisher's "*Statistical Methods for Research Workers*.")

0·1 PER CENT. POINTS OF THE DISTRIBUTION OF z.

		Values of ν_1.									
		1.	2.	3.	4.	5.	6.	8.	12.	24.	∞.
	1	6·4577	6·5612	6·5966	6·6201	6·6323	6·6405	6·6508	6·6611	6·6715	6·6819
	2	3·4531	3·4534	3·4535	3·4535	3·4535	3·4535	3·4536	3·4537	3·4536	3·4536
	3	2·5604	2·5003	2·4748	2·4603	2·4511	2·4446	2·4361	2·4272	2·4179	2·4081
	4	2·1529	2·0574	2·0143	1·9892	1·9728	1·9612	1·9459	1·9294	1·9118	1·8927
	5	1·9255	1·8002	1·7513	1·7184	1·6964	1·6808	1·6596	1·6370	1·6123	1·5845
	6	1·7849	1·6479	1·5828	1·5433	1·5177	1·4986	1·4730	1·4449	1·4134	1·3783
	7	1·6874	1·5384	1·4662	1·4221	1·3927	1·3711	1·3417	1·3090	1·2721	1·2296
	8	1·6177	1·4587	1·3809	1·3332	1·3008	1·2770	1·2443	1·2077	1·1662	1·1169
	9	1·5646	1·3982	1·3160	1·2653	1·2304	1·2047	1·1694	1·1293	1·0830	1·0279
	10	1·5232	1·3509	1·2650	1·2116	1·1748	1·1475	1·1098	1·0668	1·0165	·9557
	11	1·4900	1·3128	1·2238	1·1683	1·1297	1·1012	1·0614	1·0157	·9619	·8957
	12	1·4627	1·2814	1·1900	1·1326	1·0926	1·0628	1·0213	·9733	·9162	·8450
	13	1·4400	1·2553	1·1616	1·1026	1·0614	1·0306	·9875	·9374	·8774	·8014
	14	1·4208	1·2332	1·1376	1·0772	1·0348	1·0031	·9586	·9066	·8439	·7635
	15	1·4043	1·2141	1·1169	1·0553	1·0119	·9795	·9336	·8800	·8147	·7301
Values of ν_2.	16	1·3900	1·1976	1·0989	1·0362	·9920	·9588	·9119	·8567	·7891	·7005
	17	1·3775	1·1832	1·0832	1·0195	·9745	·9407	·8927	·8361	·7664	·6740
	18	1·3665	1·1704	1·0693	1·0047	·9590	·9246	·8757	·8178	·7462	·6502
	19	1·3567	1·1591	1·0569	·9915	·9442	·9103	·8605	·8014	·7277	·6285
	20	1·3480	1·1489	1·0458	·9798	·9329	·8974	·8469	·7867	·7115	·6086
	21	1·3401	1·1398	1·0358	·9691	·9217	·8858	·8346	·7735	·6964	·5904
	22	1·3329	1·1315	1·0268	·9595	·9116	·8753	·8234	·7612	·6828	·5738
	23	1·3264	1·1240	1·0186	·9507	·9024	·8657	·8132	·7501	·6704	·5583
	24	1·3205	1·1171	1·0111	·9427	·8939	·8569	·8038	·7400	·6589	·5440
	25	1·3151	1·1108	1·0041	·9354	·8862	·8489	·7953	·7306	·6483	·5307
	26	1·3101	1·1050	·9978	·9286	·8791	·8415	·7873	·7220	·6385	·5183
	27	1·3055	1·0997	·9920	·9223	·8725	·8346	·7800	·7140	·6294	·5066
	28	1·3013	1·0947	·9866	·9165	·8664	·8282	·7732	·7066	·6209	·4957
	29	1·2973	1·0903	·9815	·9112	·8607	·8223	·7679	·6997	·6129	·4853
	30	1·2936	1·0859	·9768	·9061	·8554	·8168	·7610	·6932	·6056	·4756
	40	1·2674	1·0552	·9435	·8701	·8174	·7771	·7184	·6463	·5513	·4016
	60	1·2413	1·0248	·9100	·8345	·7798	·7377	·6760	·5992	·4955	·3198
	∞	1·1910	·9663	·8453	·7648	·7059	·6599	·5917	·5044	·3786	0
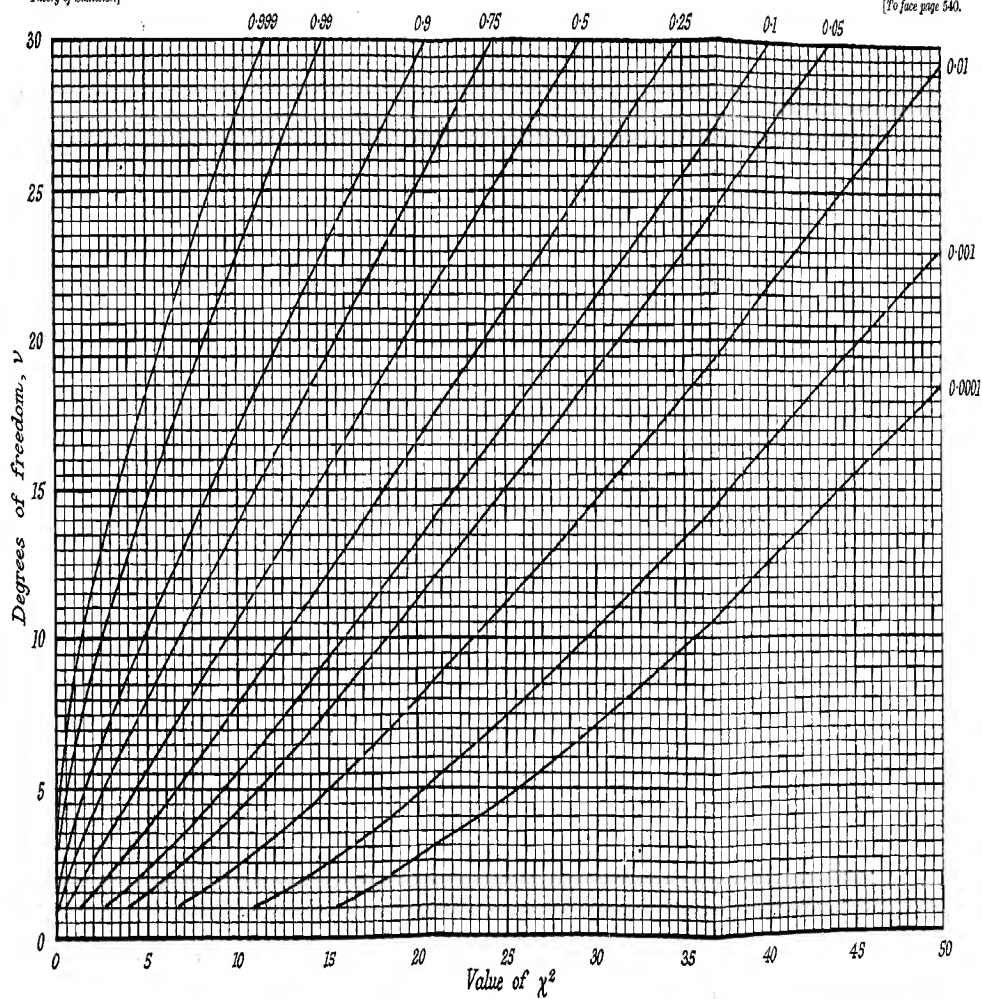
FIG. A1.—Contour Lines of the Surface $P = F(\nu, \chi^2)$.

ANSWERS

CHAPTER 1.

1.1.	N	26,287	(AB)	887
	(A)	2,308	(AC)	374
	(B)	2,853	(BC)	353
	(C)	749	(ABC)	149
1.2.	(ABC)	156	(aBC)	179
	$(AB\gamma)$	431	$(aB\gamma)$	1,249
	$(A\beta C)$	272	$(a\beta C)$	163
	$(A\beta\gamma)$	759	$(a\beta\gamma)$	20,504

1.3. The frequencies not given in the question itself are:

(a) (AB) 107 (AC) 405 (BC) 525.
(b) $(A\beta\gamma)$ 22,980 $(aB\gamma)$ 13,585 $(a\beta C)$ 96,478 $(a\beta\gamma)$ 28,868,495.

1.4. $\qquad \dfrac{(AB)}{(A\beta)} > \dfrac{(B)}{(\beta)} \qquad \therefore \quad \dfrac{(AB)}{(AB)+(A\beta)} > \dfrac{(B)}{(B)+(\beta)}$

that is $\qquad \dfrac{(AB)}{(B)} > \dfrac{(A)}{N},$ that is $\quad \dfrac{(AB)}{(B)-(AB)} > \dfrac{(A)}{N-(A)}$

that is $\qquad \dfrac{(AB)}{(aB)} > \dfrac{(A)}{(a)}$

1.5. $(AB)+(BC)-(B)$, *i.e.* the sum of the excesses of (AB) and (BC) over $(B)/2$.

1.8. 160. Take A = husband exceeding wife in first measurement, B = husband exceeding wife in second measurement, and find $(a\beta)$.

1.9. 38. If A, B, C denote passing first, second and third examinations, (C), $(a\beta C)$ and $(AB\gamma)$ are all that is necessary to answer the question. The other five frequencies (including N) are redundant.

Further, $N - (a\beta C) - (a\beta\gamma) = (A) + (B) - (ABC) - (AB\gamma)$, *i.e.* there is a linear relation between the given frequencies and the ultimate frequencies are therefore indeterminate.

1.10. 10 per cent.

CHAPTER 2.

2.1. 80/263 or 304 per thousand.
2.2. 55/85 or 65 per cent.
2.3. 32 per cent. and 30 per cent.
2.4. 117.
2.5. 108.
2.8. $p \ngtr \frac{1}{4}(1-2q)$, $p \nless \frac{1}{4}(1+2q)$, *i.e.* p must lie between 0 and $\frac{1}{4}(1-2q)$ or between $\frac{1}{4}(1+2q)$ and $\frac{1}{2}$.

2.9. As a hint, remember the condition that—

$$(BC) \nless (B) + (C) - N$$

2.10. If A, B, C denote liking chocolates, toffee or boiled sweets, $(\alpha\beta\gamma)$ is negative.

CHAPTER 3.

3.1. Deaf-mutes from childhood per million among males 222; among females 183; there is therefore positive association between deaf-mutism and male sex; if there had been no association between deaf-mutism and sex, there would have been 3176 male and 3393 female deaf-mutes.

3.2. (a) Positive association, since $(AB)_0 = 1457$.

(b) Negative association, since $294/490 = 3/5$, $380/570 = 2/3$.

(c) Independence, since $256/768 = 1/3$, $48/144 = 1/3$.

3.3.

Percentage of Plants above the Average Height.

	Parentage Crossed.	Self-fertilised.
Ipomæa purpurea . . .	86 per cent.	25 per cent.
Petunia violacea . .	79 ,,	17 ,,
Reseda lutea . . .	78 ,,	34 ,,
Reseda odorata . .	71 ,,	45 ,,
Lobelia fulgens. . .	50 ,,	35 ,,

The association is much less for the species at the end than for those at the beginning of the list.

3.4. Percentage of dark-eyed amongst the sons of dark-eyed fathers 39 per cent.

Percentage of dark-eyed amongst the sons of not dark-eyed fathers 10 per cent.

If there had been no heredity, the frequencies to the nearest unit would have been $(AB)_0$ 18, $(A\beta)_0$ 111, $(\alpha B)_0$ 121, $(\alpha\beta)_0$ 750.

3.5. Percentage of light-eyed amongst the wives of light-eyed husbands 59 per cent.

Percentage of light-eyed amongst the wives of not light-eyed husbands 53 per cent.

If there had been no association: $(AB)_0 = 298$, $(A\beta)_0 = 225$, $(\alpha B)_0 = 143$, $(\alpha\beta)_0 = 108$.

3.6. The following are the proportions of the insane per thousand in successive age-groups:—

In general population:	0·9,	2·3,	4·1,	5·7,	6·9,	7·5,	7·7, 6·8
Amongst the blind:	20·1,	16·0,	16·3,	20·7,	18·3,	17·8,	11·4, 5·3

Note the diminishing association, which is especially clear in the age-group 65–, and the negative association in the last age-group. The association coefficient gives the values below, which decrease continuously:

Association coefficient: $+0.92$, $+0.75$, $+0.61$, $+0.57$, $+0.46$, $+0.41$, $+0.20$, -0.13.

3.10. $+0.90$.

3.11. -0.70.

3.13. The frequencies are, for association:

(1)	(AB)	0
	(αB)	$(\alpha\beta)$
(2)	(AB)	$(A\beta)$
	0	$(\alpha\beta)$
(3)	(AB)	0
	0	$(\alpha\beta)$

and for disassociation:

(1)	0	$(A\beta)$
	(aB)	$(a\beta)$
(2)	(AB)	$(A\beta)$
	(aB)	0
(3)	0	$(A\beta)$
	(aB)	0

CHAPTER 4.

4.1.

$(D)/N$	$= 6.9$ per cent.	
$(AD)/(A)$	$= 45.0$	,,
$(\beta D)/(\beta)$	$= 3.6$	,,
$(A\beta D)/(A\beta)$	$= 41.2$	,,
$(BD)/(B)$	$= 42.7$	,,
$(ABD)/(AB)$	$= 51.6$	,,

$(A)/N$	$= 6.8$ per cent.	
$(AD)/(D)$	$= 44.6$	,,
$(A\beta)/(\beta)$	$= 4.7$	,,
$(A\beta D)/(\beta D)$	$= 54.9$	,,
$(AB)/(B)$	$= 29.2$	,,
$(ABD)/(BD)$	$= 35.3$	,,

The above give two legitimate comparisons. The general results are the same as for the boys, *i.e.* a very small association between development defects and dulness amongst those exhibiting nerve signs, as compared with those who do not exhibit nerve signs, or with the girls in general. As the association amongst those who do not exhibit nerve signs is quite as high as for the girls in general, the "conclusion" quoted does not seem valid.

4.2.

	(1) Per thousand.	(2) Per thousand.		(1) Per thousand.	(2) Per thousand.
$(B)/N$	3.2	7.5	$(A)/N$	0.9	4.0
$(AB)/(A)$	14.9	11.7	$(AB)/(B)$	4.0	6.3
$(BC)/(C)$	38.8	63.0	$(AC)/(C)$	6.6	18.8
$(ABC)/(AC)$	216	214	$(ABC)/(BC)$	36.8	63.8

The above give the two simplest comparisons, either of which is sufficient to show that there is a high association between blindness and mental derangement amongst the deaf-mutes as well as association in the general population; amongst the old, the association is, in fact, small for the general population, but well-marked for deaf-mutes. This result stands in direct contrast with that of Exercise 4.1, where the association between the two defects A and D was much smaller in the defective universe β than in the universe at large. As previously stated, no great reliance can be placed on the census data as to these infirmities.

4.3. If the cancer death-rates for farmers over 45 and under 45 respectively were the same as for the population at large, the rate for all farmers 15– would be 1.11. This is *slightly* less than the actual rate 1.20, but the excess would not justify the statement that "farmers were peculiarly liable to cancer." It is, in point of fact, due to the further differences of age-distribution that we have neglected, *e.g.* amongst those over 45 there are more over 55 amongst farmers than amongst the general population, and so on.

4.4. 15 per cent.

4.6. If A and B were independent in both C and γ universes, we would have (AB) equal to

$$\frac{471 \times 419}{617} + \frac{151 \times 139}{383} = 374.7$$

Actually (AB) is only 358. Therefore A and B must be disassociated in one partial universe or both.

4.9. (1) 68·1 per cent. (2) 42·5 per cent. The possible fallacy that a total association between "spending more than one's opponent" and "winning" only meant that Conservatives spent more and that Conservative principles carried the day is now avoided, and there seems no reason for declining to consider this as evidence of the effect of expenditure on election results.

4.10. The limits to y are

$$y < \tfrac{1}{2}(3x - x^2 - 1)$$
$$> \tfrac{1}{2}(x + x^2)$$

subject to the conditions $y \not> x$, $y \not< 0$, $y \not< 2x - 1$. No inference of a positive association from two negatives is possible unless x lies between the limits $0·382 \ldots, 0·618 \ldots$

4.11. The limits to y are

(1)
$$y < \tfrac{1}{2}(6x - 6x^2 - 1)$$
$$> \tfrac{1}{2}(x + 6x^2)$$

subject to conditions $y \not< 0$, $\not< 4x - 1$, $\not> x$.

An inference is only possible from positive associations of AB and AC if $x \not> \tfrac{1}{6}$; an inference is only possible from two negative associations if x lie between $0·211 \ldots$ and $0·274 \ldots$ Note that x cannot exceed $\tfrac{1}{3}$.

(2)
$$y < \tfrac{1}{2}(6x - 3x^2 - 1)$$
$$> \tfrac{1}{2}(2x + 3x^2)$$

subject to conditions $y \not< 0$, $\not< 5x - 1$, $\not> x$.

No inference is possible from positive associations of AB and BC.

An inference is only possible from negative associations if x lie between $0·183 \ldots$ and $0·215 \ldots$ Note that x cannot exceed $\tfrac{1}{4}$.

(3)
$$y < \tfrac{1}{2}(6x - 2x^2 - 1)$$
$$> \tfrac{1}{2}(3x + 2x^2)$$

subject to the conditions $y \not< 0$, $\not< 5x - 1$, $\not> x$.

As in (2), no inference is possible from positive associations of AC and BC; an inference is possible from negative associations if x lie between $0·177 \ldots$ and $0·224 \ldots$ Note that x cannot exceed $\tfrac{1}{4}$.

CHAPTER 5.

5.1. A, 0·68; B, 0·36.

5.2. $C = 0·02$, $T = 0·01$.

5.4. The table is not isotropic as it stands. It becomes positively so if the columns are arranged in the order A_1, A_3, A_5, A_4, A_2, and the rows in order (from top to bottom) B_3, B_2, B_1.

5.5. $C = 0·05$, $T = 0·03$.

5.7. $C = 0·40$. For a large number such as 1000 this is probably significant, *i.e.* not due to fluctuations of sampling. From inspection of the tables the contingency is positive, *i.e.* this evidence would suggest that persons tended on the whole to prefer music of their own nationality. But there are exceptions, *e.g.* the English.

In any case these data are purely imaginary, and it is not suggested that they reflect in any way the true state of affairs.

5.8. $C = 0·23$, $T = 0·17$, suggestive of slight association.

5.10. $C = 0·10$.

CHAPTER 6.

6.1. 1200, 200.

6.2. 270, 40.

6.3. 95·75.

6.4. 216·5.

6.5. (a) J-shaped; (b) U-shaped; (c) single-humped moderately asymmetrical; (d) J-shaped in all three cases.

CHAPTER 7.

7.2. 14·58.

7.3. Mean, 156·73 lb. Median, 154·67 lb. Mode (approx.), 150·6 lb. (Note that the mean and the median should be taken to a place of decimals further than is desired for the mode; the true mode, found by fitting a theoretical frequency curve, is 151·1 lb.)

7.4. Mean, 0·6330. Median, 0·6391. Mode (approx.), 0·651. (True mode is 0·653.)

7.5. About £3250.

7.6. Mean $=\dfrac{n+1}{2}$.

7.7. (1) 82·75, (2) 81·78, (3) 80·25, (4) 80·25.

7.8. Arithmetic mean $=\dfrac{1}{n+1}(2^{n+1}-1)$

Geometric mean $=2^{\frac{n}{2}}$.

Harmonic mean $=\dfrac{n+1}{2\left(1-\dfrac{1}{2^{n+1}}\right)}$.

7.9. Mean $=np$. If the terms of the given binomial series are multiplied by 0, 1, 2, . . ., note that the resulting series is also a binomial when a common factor is removed. (A full proof is given in Chapter 10.)

7.11. (1) 921,507, (2) 916,963.

7.12. For N.M. specials, 15s. 1d. per 120; for ordinaries, 12s. 9d. per 120.

CHAPTER 8.

8.2. Standard deviation 21·3 lb. Mean deviation 16·4 lb. Lower quartile 142·5, upper quartile 168·4; whence $Q=12·95$. Ratios: m.d./s.d. $=0·77$, Q/s.d. $=0·61$.

8.3. Median $=£3250$, upper quartile $=£5000$, 9th decile $=£8600$ approximately.

8.4. $Q_1=24·13$ years. Median $=27·29$ years. $Q_3=32·19$ years. $Q=4·03$ years.

8.5. 2·872.

8.6. This proposition is equivalent to the one that the square of the mean of a set of positive numbers is less than the mean of the squares. This is proved in most text-books on Algebra.

8.8. (1) $M=73·2$, $\sigma=17·3$; (2) $M=73·2$, $\sigma=17·5$; (3) $M=73·2$, $\sigma=18·0$. (Note that while the mean is unaffected in the first place of decimals, the standard deviation is higher the coarser the grouping.)

8.9. England, $\sigma=2·55$; Scotland, $\sigma=2·48$; Wales, $\sigma=2·33$; Ireland, $\sigma=2·15$ inches. For the weight distribution $\sigma=21·14$ lb.

8.10. $\sqrt{npq}$. The proof is given in Chapter 10.

35

8.11. The assumption that observations are evenly distributed over the intervals does not affect the sum of deviations, except for the interval in which the mean or median lies; for that interval the sum is $n_2(0\cdot25+d^2)$, hence the entire correction is

$$d(n_1-n_3)+n_2(0\cdot25+d^2)$$

In this expression d is, of course, expressed as a fraction of the class-interval, and is given its proper sign.

8.14. 3·80, 3·65, 3·53, 3·20.

CHAPTER 9.

9.1. In class-intervals of 10 lb.

$\mu_2=4\cdot470,\ \mu_3=6\cdot927,\ \mu_4=89\cdot119;\ \beta_1=0\cdot537,\ \beta_2=4\cdot461.$

Curve leptokurtic.

9.2. 0·06, 0·29, 0·27.

9.3. $\mu_2=11\cdot375,\ \mu_3=12\cdot705,\ \mu_4=428\cdot708$, in class-intervals of 1 gallon.

$\beta_1=0\cdot110,\ \beta_2=3\cdot313.$

Measures of skewness are 0·027, 0·14, 0·15. The second is obtained by approximating to the mode in the manner of 7.26.

9.4. Before corrections, $\mu_2=7\cdot301,\ \mu_3=0\cdot166,\ \mu_4=163\cdot465$;

After corrections, $\mu_2=6\cdot551,\ \mu_3=0\cdot166,\ \mu_4=132\cdot975.$

Note that the small negative μ_3 in the finer grouping becomes positive in the coarser grouping.

9.5. $\mu_3=npq(q-p).$

$\mu_4=3p^2q^2n^2+pqn(1-6pq).$

9.6. About the mean, $\mu_2=14\cdot75,\ \mu_3=39\cdot75,\ \mu_4=142\cdot3125.$

About the origin, $\mu_2'=21,\ \mu_3'=166,\ \mu_4'=1132.$

9.8. This proposition is equivalent to that of Exercise 8.6. For U-shaped universes $\beta_2<2.$

9.9. $\lambda_2=7\cdot057,\ \lambda_3=36\cdot152,\ \lambda_4=259\cdot335.$

CHAPTER 10.

10.1. 27·31 per cent.

10.2. Expected frequencies are: 1, 12, 66, 220, 495, 792, 924, 792, 495, 220, 66, 12, 1.

Expected mean $=6$; expected $\sigma=1\cdot732.$

Actual mean $=6\cdot139$; actual $\sigma=1\cdot712.$

10.3. $y=\dfrac{4096}{1\cdot712\sqrt{2\pi}}e^{-\frac{1}{2(1\cdot712)^2}(x-6\cdot139)^2}$

Expected frequencies, to nearest unit, are: 2, 11, 51, 178, 438, 765, 951, 841, 529, 236, 75, 17, 3, totalling 4097; (these are obtained by simple interpolation in Appendix Table 1).

10.4. 1?.

10.5. If p is the expectation of getting an even number,

$$^{10}C_5p^5q^5=2\times{}^{10}C_4p^4q^6$$

Hence, $p=\frac{2}{3}$, and the number of times is $10{,}000(\frac{2}{3})^{10}=$ once.

10.8. The frequency of r successes is greater than that of $r-1$ so long as $r<np+p$; if np is an integer, $r=np$ gives the greatest term and also the mean.

10.9. This follows at once from a consideration of the Galton-Pearson apparatus.

10.10.

Binomial.	Normal Curve.
1	1·7
10	10·5
45	42·7
120	116·1
210	211·5
252	258·4
210	211·5
etc.	etc.

10.11. Mean 74·3, standard deviation 3·23.

10.12. About zero mean the deciles are: 0, 0·2533, 0·5244, 0·8416, 1·2816, and the corresponding negative values.

10.13. $y = \dfrac{8585}{2 \cdot 57 \sqrt{2\pi}} e^{-\frac{1}{2(2 \cdot 57)^2}(x - 67 \cdot 46)^2}$

Calculated mean and quartile deviations, 2·05 and 1·73 (observed, 2·02 and 1·75). These figures are in units of one inch.

10.14. Calculated mean and quartile deviations (years), 6·37 and 5·38 (observed, 5·44 and 4·03).

10.15. 18.

10.16. $\sigma = 2·267$ (uncorrected).

Theoretical frequencies, 2, 5, 11, 20, 29, 35, 35, etc.

10.17. Theoretical frequencies, 336·5, 397·1, 234·6, 92·5, 27·3, 6·5, 1·3, 0·2.

10.18. $\lambda_2 = 1·362$, $\lambda_3 = 1·766$, $\lambda_4 = 2·510$.

CHAPTER 11.

11.1. $\sigma_x = 1·414$, $\sigma_y = 2·280$, $r = +0·81$.

$X = 0·5Y + 0·5$, $Y = 1·3X + 1·1$.

11.2. r (between X and Y) $= -0·66$; between Y and $Z = 0·60$; between Z and $X = -0·13$.

11.4. $r = +0·96$.

11.5. (1) $-0·41$, (2) $+0·40$.

CHAPTER 12.

12.3. From equations (12.11) and (12.12) replace σ_1 and σ_2 by S_1 and S_2 in equation (12.10). Regarding this as an equation for r, note that r^2 is a maximum when $\tan 2\theta$ is infinite, or $\theta = 45°$.

12.4. In fig. 12.1 suppose every horizontal array to be given a slide to the right until its mean lies on the vertical axis through the mean of the whole distribution; then suppose the ellipses to be squeezed in the direction of this vertical axis until they become circles. The original quadrant has now become a sector with an angle between one and two right angles, and the question is solved on determining its magnitude.

12.5. The ellipse is a horizontal section of the surface. Its equation is

$$\frac{x^2}{\sigma_1^2} - \frac{2rxy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} = 1 - r^2$$

and the standard deviations of sections are the square roots of the lengths of radii vectors of the ellipse.

12.6. The maximum and minimum s.d.'s are given by the principal axes, which leads to equations (12.11) and (12.12).

For an intermediate value there are two radii vectors and hence two sections.

12.7. a and b must be negative, and $ab - h^2 \geqslant 0$.

$$\sigma_1{}^2 = -\tfrac{1}{2}\frac{b}{ab-h^2}, \qquad \sigma_2{}^2 = -\tfrac{1}{2}\frac{a}{ab-h^2}$$

$$r = \frac{h}{\sqrt{ab}}$$

CHAPTER 13.

13.1. $\eta_{xy} = 0 \cdot 242$, $\eta_{yx} = 0 \cdot 266$.

13.2. $\eta_{xy} = 0 \cdot 82$, $\eta_{yx} = 0 \cdot 80$.

13.3. $\rho = +0 \cdot 79$.

13.4. If the judges be denoted by 1, 2, 3,

$$\rho_{12} = -0 \cdot 21, \qquad \rho_{23} = -0 \cdot 30, \qquad \rho_{13} = +0 \cdot 64$$

This suggests that judges 1 and 3 have tastes in common, but neither has much in common with judge 2.

13.5. $Q = 2/3$.

13.6. $Q = 0 \cdot 77$.

13.8. $r = +0 \cdot 83$.

13.9. $r = +0 \cdot 22$, 11,868 entries.

CHAPTER 14.

14.1. $r_{12.3} = +0 \cdot 759$, $r_{13.2} = +0 \cdot 097$, $r_{23.1} = -0 \cdot 436$.
$\sigma_{1.23} = 2 \cdot 64$, $\sigma_{2.13} = 0 \cdot 594$, $\sigma_{3.12} = 70 \cdot 1$.
$X_1 = 9 \cdot 31 + 3 \cdot 37 X_2 + 0 \cdot 00364 X_3$.

14.2. $R_{1(23)} = 0 \cdot 80$, $R_{2(31)} = 0 \cdot 84$, $R_{3(12)} = 0 \cdot 57$.

14.3. $r_{12.34} = +0 \cdot 680$, $r_{13.24} = +0 \cdot 803$, $r_{14.23} = +0 \cdot 397$.
$r_{23.14} = -0 \cdot 433$, $r_{24.13} = -0 \cdot 553$, $r_{34.12} = -0 \cdot 149$.
$\sigma_{1.234} = 9 \cdot 17$, $\sigma_{2.134} = 49 \cdot 2$, $\sigma_{3.124} = 12 \cdot 5$, $\sigma_{4.123} = 105 \cdot 4$.
$X_1 = 53 + 0 \cdot 127 X_2 + 0 \cdot 587 X_3 + 0 \cdot 0345 X_4$.

14.4. $R_{1(23)} = 0 \cdot 64$, $R_{1(234)} = 0 \cdot 72$.

14.5. $(X_1 - 19 \cdot 9) = 4 \cdot 51 (X_2 - 49 \cdot 2) - 0 \cdot 88 (X_3 - 30 \cdot 2)$
$$- 0 \cdot 072 (X_4 - 4814) + 0 \cdot 63 (X_5 - 41 \cdot 6).$$

$$r_{15.3} = -0 \cdot 03.$$
$$r_{15.4} = +0 \cdot 25.$$
$$r_{15.34} = +0 \cdot 23.$$
$$R_{1(2345)} = 0 \cdot 77.$$

14.7. Number of order $s = n \times {}^{n-1}C_s$
Total number $= n\{2^{n-1} - 1\}$
This includes coefficients of type $R_{1(2)}$ and counts $R_{1(2)}$ as different from $R_{2(1)}$.

14.8. The correlation of the pth order is $r/(1 + pr)$. Hence if r be negative, the correlation of order $n - 2$ cannot be numerically greater than unity and r cannot exceed (numerically) $1/(n - 1)$.

14.9. $r_{12.3} = -1$, $r_{13.2} = r_{23.1} = +1$.

14.10. $r_{12.3} = r_{13.2} = r_{23.1} = -1$.

CHAPTER 16.

16.1. Estimated true standard deviation $6 \cdot 91$; standard deviation of fluctuations of sampling $9 \cdot 38$. (The latter, which can be independently calculated, is too low, and the former consequently probably too high. *Cf.* 19.30.)

16.2. $0 \cdot 43$.

16.3. 58 per cent.

16.4. $\sigma_2{}^2 / \sqrt{(\sigma_1{}^2 + \sigma_2{}^2)(\sigma_2{}^2 + \sigma_3{}^2)}$

16.5. $\dfrac{a\sigma_1}{\sqrt{a^2\sigma_1{}^2 + b^2\sigma_2{}^2}}$

16.6. 0·29.

16.7. $r_{13} = \dfrac{1}{2ab\sigma_1\sigma_3}(-a^2\sigma_1{}^2 - b^2\sigma_2{}^2 + c^2\sigma_3{}^2)$

The others may be written down from symmetry.

16.8. (1) No effect at all. (2) If the mean value of the errors in variables is d, and in the weights e, the value found for the weighted mean is:

$$\text{The true value} + d - r \cdot \sigma_x \cdot \sigma_w \frac{e}{\bar{w}(\bar{w} + e)}$$

If r is small, d is the important term, and hence errors in the quantities are usually of more importance than errors in the weights. If r become considerable, errors in the weights may be of consequence, but it does not seem probable that the second term would become the most important in practical cases.

16.9. $r = +0·036.$

CHAPTER 17.

17.1. Line: $Y = 2·58 + 1·13(X - 2)$
       Quadratic: $Y = 1·48 + 1·13(X - 2) + 0·55(X - 2)^2$
       Cubic: $Y = 1·48 + 0·025(X - 2) + 0·55(X - 2)^2 + 0·325(X - 2)^3$
       Sums of squares of residuals: 5·819, 1·584, 0·063.

17.2. If Y is the average number of children for the duration X to $X + 1$ years :

Line: $Y = 3·814 + 0·887\left(\dfrac{X}{5} - 3\right)$

Quadratic: $Y = 4·351 + 0·887\left(\dfrac{X}{5} - 3\right) - 0·134\left(\dfrac{X}{5} - 3\right)^2$

Cubic: $Y = 4·351 + 0·365\left(\dfrac{X}{5} - 3\right) - 0·134\left(\dfrac{X}{5} - 3\right)^2 - 0·00361\left(\dfrac{X}{5} - 3\right)^3$

For $X = 17$ the three values are 4·17, 4·68, 4·69.

17.3. $\gamma = 1·42.$

17.4. $X = $ Gross output per £100 labour, $Y = $ gross output.
       $Y = 48·33 + 0·2375X - 0·00005546X^2$

CHAPTER 19.

19.1. Theo. $M = 6$, $\sigma = 1·732$: Actual $M = 6·116$, $\sigma = 1·732$.

19.2. (a) Theo. $M = 2·5$, $\sigma = 1·118$: Actual $M = 2·48$, $\sigma = 1·14$.
     (b)   ,, $M = 3$, $\sigma = 1·225$:   ,, $M = 2·97$, $\sigma = 1·26$.
     (c)   ,, $M = 3·5$, $\sigma = 1·323$:   ,, $M = 3·47$, $\sigma = 1·40$.

19.3. The standard deviation of the proportion is 0·00179, and the actual divergence is 5·4 times this, and therefore almost certainly significant.

19.4. The standard deviation of the number drawn is 32, and the actual difference from expectation 18. There is no significance.

19.5. Difference from expectation 7·5; standard error 10·0. The difference might therefore occur frequently as a fluctuation of sampling.

19.6. Standard error of proportion of bad eggs = 1·6536 per cent. A range of three times this gives range of 7·5 per cent. to 17·5 per cent. approximately.

19.7. The test can be applied either by the formulæ of Case 2 (19.28) or those of Case 3 (19.29). Case 2 is taken as the simplest.

$$(AB)/(B) = 70 \cdot 1 \text{ per cent.}; \quad (A\beta)/(\beta) = 64 \cdot 3 \text{ per cent.}$$

Difference 5·8 per cent. $(A)/N = 67 \cdot 6$ per cent. and thence $\epsilon_{12} = 3 \cdot 40$ per cent. The actual difference is 1·7 times this and might, rather infrequently, occur as a fluctuation of sampling.

19.9. Difference of proportions $= \frac{1}{10}$, $\epsilon_{12} = 0 \cdot 033$. Difference significant. Similar conclusions follow if the formulæ of Case 3 (19.29) are applied.

19.10. Proportion $= 36$ per cent. Limits $32 \cdot 4 - 39 \cdot 6$ per cent. The sampling is almost certainly not simple. Possible causes are: (a) nature of subject-matter might require words of certain type, e.g. scientific words probably would not be Anglo-Saxon; (b) the occurrence of one word influences the occurrence of the next.

19.11. If there are f_1 samples of n_1 individuals each, f_2 of n_2, etc.,

$$Ns^2 = pq\left(\frac{f_1}{n_1} + \frac{f_2}{n_2} + \cdots\right)$$

$$s^2 = \frac{pq}{H}$$

19.12. Standard error of expected proportion $= 23 \cdot 05$ per cent.
Standard deviation of actual distribution $= 23 \cdot 09$ per cent.

19.13. Standard deviation of simple sampling 23·0 per cent. The actual standard deviation does not, therefore, seem to indicate any real variation, but only fluctuations of sampling.

19.14. $s_0 = 7 \cdot 02$, and $\sigma_p = 2 \cdot 5$ units.

19.15. $\sigma^2 = npq$ as if the chance of success were p in all cases (but the mean is $n/2$, not pn).

19.16. Mean number of deaths per annum $= \sigma_0^2 = 680$,

$$\sigma^2 = 566{,}582 \qquad r = 0 \cdot 000029.$$

CHAPTER 20.

20.1. $P = 0 \cdot 1773$.

20.2. $P = 0 \cdot 9595$.

20.3. Median: Estimated frequency $= 1554$. Standard error 0·28 lb.
Lower Q: frequency 1472. Standard error 0·26 lb.
Upper Q: frequency 1116. Standard error 0·34 lb.

20.4. 0·18 lb.

20.5. 0·24 lb., 14 per cent. less than the s.e. of the median.

20.6. Estimated frequencies: $Q_1 = 67{,}548$, $Mi = 63{,}152$, $Q_3 = 30{,}488$.
Standard errors (years) 0·011, 0·013, 0·023.

20.7. Standard error of mean $= 0 \cdot 015$ years.

20.8. Standard error of quartiles 0·020 years.

20.9. $\dfrac{\sigma}{\sqrt{n}} \times 1 \cdot 34270$.

20.10. $\epsilon_{12} = 1 \cdot 36$ shillings. Difference of means 2 shillings. Difference hardly suggestive of real effect.

20.12. Yes, one might, because the results on farms in successive years are correlated.

20.13. Mean $= 5 \cdot 613$; s.e. of mean 0·10.
Median $= 8 \cdot 128$; s.e. of median 0·21.

20.14. $P = 0 \cdot 309$.

20.15. £450,000; £1,350,000.

20.16. 0·72 inch.

CHAPTER 21.

21.1. Standard error $=0\cdot223$ lb.
On basis of normal distribution $=0\cdot170$ lb.

21.2. $0\cdot011$, $0\cdot014$.

21.3. S.e. of s.d. $=0\cdot707\dfrac{\sigma}{\sqrt{n}}$

S.e. of Q $=0\cdot787\dfrac{\sigma}{\sqrt{n}}$

21.4. Difference of s.d.'s $0\cdot2$. On the assumption of normality $\epsilon_{12}=0\cdot088$. Difference might therefore arise, rather infrequently, as sampling fluctuation.

21.5. $r= -0\cdot008$ for height distribution, $r= +0\cdot71$ for marriage distribution.

21.6. $\sigma_{\lambda_1}^2 = \dfrac{\sigma^2}{n}$

$\sigma_{\lambda_2}^2 = \dfrac{\mu_4 - \mu_2}{n} = \dfrac{2\sigma^4}{n}$ for normal curve.

$\sigma_{\lambda_3}^2 = \dfrac{\mu_4 - \mu_3{}^2 - 6\mu_4\mu_2 + 9\mu_2{}^3}{n} = \dfrac{6\sigma^6}{n}$ for normal curve.

$\sigma_{\lambda_4}^2 = \dfrac{1}{n}\{36\mu_2{}^2(\mu_4 - \mu_2{}^2) + (\mu_8 - \mu_4{}^2 - 8\mu_5\mu_3)$
$\qquad\qquad\qquad + 16\mu_2\mu_3{}^2 - 12\mu_2(\mu_6 - \mu_2\mu_4 - 4\mu_3{}^2)\}$

$\quad = \dfrac{24\sigma^8}{n}$ for normal curve.

21.7. For the 6th and lower moments.

21.9. Standard errors are $0\cdot0176$, $0\cdot0158$, $0\cdot0263$, and results might all have arisen from an uncorrelated universe; if the universe were actually uncorrelated, the standard errors would be the same to the number of places given, owing to the smallness of r.

21.10. Standard errors $0\cdot0758$, $0\cdot1308$, $0\cdot0850$, and the correlations are all significant.

CHAPTER 22.

22.1. $\chi^2 = 5\cdot811$, $\nu = 7$, $P = 0\cdot56$.

22.3. $\chi^2 = 4\cdot3$, $\nu = 9$, $P = 0\cdot89$. The hypothesis seems reasonable.

22.5. $\chi^2 = 27\cdot94$, $\nu = 4$, $P = 0\cdot000012$. The association is significant.

22.6. $\chi^2 = 0\cdot7080$, $\nu = 1$, $P = 0\cdot400$. The divergences from expectation may well have arisen by sampling fluctuations.

22.7. Use the result that for large n, χ^2 is distributed approximately normally.

22.8. $\chi^2 = 27\cdot68$, $\nu = 4$, $P = 0\cdot00001$. The data are very suggestive of association.

22.11. $\chi^2 = 13\cdot15$, $\nu = 2$, $P = 0\cdot0014$. This is rather low and we suspect the sampling to be non-random.

22.12. $\chi^2 = 9\cdot993$, $\nu = 3$, $P = 0\cdot018$. Not a very good fit. (In this Exercise the last four frequencies have been grouped together and ν reduced by unity to allow for the estimation of the mean of the Poisson distribution.)

22.14. $\chi^2 = 0\cdot4700$, $\nu = 3$, $P = 0\cdot943$ (by direct calculation).

CHAPTER 23.

23.1. $t = -0.664$, $\nu = 9$, $P = 0.738$.

The probability that we should get a value of t greater *in absolute value* is 0·524.

23.2. The differences in the returns, including cost of manure, have mean $= 1$, $_c\sigma_s^2 = 1.375$, $t = 1.907$, $\nu = 4$, $P = 0.935$.   Assuming that distribution of differences is normal, a greater value would arise about 65 times in 1000.   There is some reason for supposing that the increased returns on the better manured plot are real, and that it would therefore pay to continue the more expensive dressing.

23.3. Applying the t test for two samples,

$$t = 0.0991, \qquad \nu = 14, \qquad P = 0.54$$

There is nothing in this test to suggest that universes were unlike as regards height.

23.4. $z = 0.1761$, $\nu_1 = 9$, $\nu_2 = 5$.   The difference of standard deviations is not significant.   Coupled with Exercise 23.3, we conclude that there is no ground for supposing the two universes different as regards height.

23.5. Applying the t test for two samples,

$$t = 2.683, \qquad \nu = 4, \qquad P = 0.972$$

The difference of means is likely to be significant, which supports the suggestion.

23.6. $z = \frac{1}{2} \log_e \frac{1+r}{1-r} = -0.549$ $\qquad \sigma = \frac{1}{\sqrt{12}} = 0.2887$

The observed deviation is suggestive, but not decisive.

23.8. $P = 0.0048$.   For the standard error formula $P = 0.0000078$.

23.9. All significant.

23.10. All significant.

23.12. Significantly non-linear.

CHAPTER 24.

24.1. 0·93877, 0·93823, 0·93822.

24.2. 0·823632, 0·818050, 0·817939.   The inclusion of the third difference affects only the fourth place by a single unit, so we can probably trust the answer to four figures.

24.3. Using logarithmic interpolation, the successive approximations are : 0·11200, 0·10044, 0·09963.   Second difference interpolation using the last three data only gives 0·09859.   It looks as if we could trust the figure as about 0·100 or 0·099.

24.4. 4195, 4443, 4724, 5036, 5380.

24.7. 11·388 approximately.

24.8. Median 4·8924, 4·8869.   First decile 1·9474, 1·9572.   Ninth decile 8·4286, 8·3733.   As we would probably state such figures only to two decimal places, the median would not be appreciably affected by taking second differences into account, but the deciles would be slightly corrected.

24.9. Maximum at 1·336, or day 40, 25th July, value 63·7.

Minimum at 1·184, or day 35·5, 20th–21st January, value 38·0.

These estimates are very poor.   The maximum is actually 63·4 on 15th–17th July, and the minimum 37·9 on 8th–12th January.

INDEX.